THE DEVELOPMENT OF A MODEL FOR BEHAVIORAL OBSERVATION PERFORMANCE OF INSTRUCTIONAL PERSONNEL

A Dissertation

Presented to

The School of Graduate Studies

Drake University

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Education

---

by

Raymond L. Feltner

May 1983

THE DEVELOPMENT OF A MODEL FOR BEHAVIORAL OBSERVATION PERFORMANCE OF INSTRUCTIONAL PERSONNEL

by

Raymond L. Feltner

Approved by Committee:

Dr. Richard H. Lampshire, Chair

Dr. Harvey A. Martens

Dr. Paul C. Vance

Dr. Earle L. Canfield
Dean of the School of Graduate Studies

THE DEVELOPMENT OF A MODEL FOR BEHAVIORAL OBSERVATION PERFORMANCE OF INSTRUCTIONAL PERSONNEL

An abstract of a Dissertation by
Raymond L. Feltner
May 1983
Drake University
Advisor: Richard H. Lampshire

The problem. The purpose of the project was to develop a performance evaluation model that would measure teaching behaviors of instructional personnel on a systematic and objective basis. The model was intended to reduce subjectivity and bias and yield high reliability and agreement among raters.

Procedure. There were two urban and one rural school districts that participated in field testing the model for evaluating instructional personnel. After the instruments were developed, each district field tested the model by actual classroom observation of teaching activities.

Twenty-two teachers from kindergarten through twelfth grade and special education were selected by school principals to be observed. Each observation for field testing and collecting data for the model was conducted by observers in teams of three.

After the data were returned, they were analyzed for correlation of coefficient to determine reliability between raters. A percentage of agreement among raters was also determined from the data.

Findings. The analysis provided inter-rater reliability of .90 for sixteen of the nineteen classroom observations. Mean inter-rater agreement of 85 percent was found on sixteen of the nineteen observations. Results of the three field sites were quite similar. Indications were the size of school district or grade taught did not influence the results.

Conclusions. The results of the project provide support for the statement that a model can be developed for evaluating instructional personnel with high reliability and inter-rater agreement. The conclusions can be drawn: (1) Instructional personnel can be objectively and reliably observed; (2) Principals can be trained to observe and accurately identify teaching behaviors; (3) Previous studies of behavior observation methods have been supported by this project.

Recommendations. Recommendations included: (1) School districts planning to replicate the model emphasize the procedures for training; (2) Instructional personnel be included in any design or modification of a performance evaluation model; (3) A five point, rather than a three point, scalogram be developed for each behavior scale; (4) Greater representation of the performance evaluation be obtained through reviews by educational administration preparation programs.

Table of Contents

Chapter Page

Tables

Table Page

Table Page

Table | Page

Figures

CHAPTER ONE

Introduction

Performance evaluation of instructional personnel is frequently described in the literature as an activity of little value to either the school district or the instructional staff. Educators are recognizing that currently designed instruments to measure teacher performance are inappropriate for assessing either teacher performance or the quality of that performance.[1] A minimal expectation of a performance evaluation model is that it provide objective and reliable feedback to instructional personnel.

School districts make little use of performance evaluation results for either selection, placement or training purposes.[2] According to Thomas, the reason is primarily due to the design and subjectivity of currently used procedures.[3] Evaluation instruments need to be developed containing procedures that are clearly defined and that record objective

---

[1]Herbert J. Walberg, Evaluating Educational Performance: A Sourcebook of Methods, Instruments and Examples (Berkeley, California: McCutchan Publishers, 1974), pp. 1-9.

[2]Donald M. Thomas, Performance Evaluation of Educational Personnel (Bloomington, Indiana: Phi Delta Kappa Educational Foundation, 1979), pp. 3-12.

[3]Ibid.

results. The process of evaluation for local school districts to follow is scarce in the literature. The intent of this project was to offer a model for implementing objective performance evaluation of instructional behaviors.

Rationale

Performance evaluation of instructional personnel is a complex process, but a process needed in every school district. From early in the history of education various methods have been used to evaluate instructional personnel. According to Levin, research provides little support for current practices in teacher evaluation.[1] Although a wide variety of performance appraisal instruments is available, specific school district objectives will be difficult to meet unless the appraisal instrument accurately measures instructional activities of the person being rated. Behavioral based rating models are a fairly new development that would seem to greatly increase the effectiveness of teacher evaluations and to meet the school districts objectives and goals.

The purpose of this project was to develop a model that school districts could emulate for writing evaluation scales related directly to the districts' objectives and goals. Thus, a school district's performance appraisal procedures could be established according to their own established or

---

[1]Benjy Levin, "Teacher Evaluation: A Review of Research," Educational Leadership, 37 (December 1979), 240-45.

developed goals rather than having to adopt or adapt commercial instruments that do not relate to a specific educational community. The development of a model for evaluating instructional personnel was predicated upon the concept that instruction is a professional and significant activity requiring cooperative and competent teachers working toward mutual goals.

Methods That Have Biases

Performance evaluation for instructional improvement receives a generous share in the educational literature. Over the years a wide variety of performance evaluation methods in education have been developed and used.

Early in the history of education, teachers were evaluated on the basis of traits and attributes. Educators and researchers began to question this method. As a result, new programs for performance evaluation were implemented. These new performance evaluation methods concentrated more on relationships between teacher and others (student, parent, associates, etc.) and skills that could be demonstrated by educators. Disagreement continued to exist as to whether standard, specific instructional skills can be identified and used for every school district.

Research does indicate that effective teachers tend to have certain competencies. This has been demonstrated by

the work of Rosenshine and Furst.[1] Even they are cautious to point out that effective teaching is relative to a "cluster" of competencies and not to individual skills. Again, as with traits, certain skills, although representing a segment of the total effective teacher, are difficult to isolate and measure effectively. According to Roy, "there is no adequate reason to believe that educators can be evaluated by marking a rating scale that contains a long list of skills and competencies."[2]

In recent years efforts have been directed to product evaluation methods, such as evaluation on student achievement. Borg contends this particular evaluation process does not give adequate consideration to the many variables that affect the product.[3] This method has been criticized because a teacher might be more successful with one group of students than with another group. Rosenshine reviewed several studies of stability with the conclusion that stability was low.[4] The technique seems to be used infrequently and little

---

[1]Barak Rosenshine and Norma Furst, "The Use of Direct Observation to Study Teaching," in Second Handbook of Research on Teaching, ed. Robert M. W. Travers (Chicago: Rand McNally, 1973), pp. 129-83.

[2]Joseph J. Roy, "Teacher Evaluation in an Era of Educational Change," The Clearinghouse, 52 (February 1979), 275.

[3]Walter R. Borg, Applying Educational Research (New York: Longman, 1981), pp. 90-101.

[4]Rosenshine and Furst, pp. 129-83.

research has been conducted.

Evaluation of teachers by students also became a popular trend in the sixties, especially at the university level. This approach has remained in fairly wide use at the university level, but is less common in the public schools. Therefore, most of the research that has been conducted has involved college students. The major concern with this method has been the question of reliability and bias.

Methods Having the Most Potential

Current performance evaluation programs are being based on performance standards rather than traits, skills, and product analysis. Direct observation is the more accepted method for collecting the data. According to Van Dalen, the effectiveness of rating scales depends in part upon the qualifications of the raters with individuals often checking scale choices on the basis of inadequate evidence.[1] Levin states "the use of techniques that have greater promise for providing objective data, such as observation, is yet uncommon."[2]

The literature on teacher evaluation is enormous. However, when the criteria of reliability, written performance criteria and objective-systematic observed behaviors are

---

[1]Deobold B. Van Dalen, Understanding Educational Research (New York: McGraw-Hill, 1973), pp. 36-61.

[2]Levin, p. 244.

specified, the amount reduces sharply. Systematic observation of instructional personnel involves the use of an instrument to guide the observer in terms of the behavior to be observed. Levin states that the use of such instruments is quite rare. Researchers, however, use the observation method extensively in the study of classroom behavior. Borich finds that observation has two important advantages over the more judgmental approaches.[1] The behaviors observed are both pre-specified and classroom-based. Enns reports that teacher evaluation has been viewed with pessimism in the past, but is now viewed with increasing optimism because of the aspect of classroom observation.[2] Herman believes that the use of clearly stated behavioral objectives subject to measurement, coupled with observational instruments hold the greatest promise for creative and objective evaluative systems.[3] Rosenshine and Furst support this on the premise that the characteristics of effective teaching are observable by raters.[4] Van Dalen is even more crisp in

---

[1]Gary D. Borich and Susan K. Madden, Evaluating Classroom Instruction: A Source Book of Instruments (Reading, Massachusetts: Addison-Wesley, 1977), pp. 1-13.

[2]T. Enns, "Rating Teacher Effectiveness: The Functions of the Principal," The Journal of Educational Administration, 3 (March 1965), 81-95.

[3]Jerry J. Herman, Developing an Effective School Staff Evaluation Program (West Nyack, N.Y.: Parker Publishing, 1973), pp. 23-30.

[4]Rosenshine and Furst, pp. 129-83.

his support with the statement "observation is fundamental in research, for it produces basic elements of science--facts."[1] He also notes that observation instruments are reputable since they list items (carefully defined, observable factors) relevant to the situation. Thus, the clearer the definition of the units to be observed and evaluated, the fewer the inferences required of the observer.

Theory should guide the selection of any method. The methodology chosen can affect the kind of information that will be obtained for evaluating instructional personnel. The model then becomes a methodological tool to guide and focus on observable instructional activities.

Statement of the Problem

The purpose of this project was to develop a performance evaluation model that measures the behaviors of instructional personnel on a systematic, objective basis. The model was to reduce subjectivity and bias with high inter-rater reliability, having at least an 85 percent agreement between raters.

Limitations

The limitations of the project were: (1) the school population participating in the three field demonstration sites, (2) the participating teachers and school districts

---

[1]Van Dalen, p. 36.

were not randomly sampled.

In spite of these, it is anticipated by the developer of the model that the project results can be generalized to any public or private school district. The results would be limited to school districts' (1) general education instructional personnel in grades K-12; (2) special education instructional personnel at the elementary and secondary levels.

Definitions

Performance evaluation: The systematic evaluation of individual instructional performance.

Inter-rater reliability: The consistency or agreement between two or more independently derived observations recorded on the same instrument.

Behavior: A particular instructional activity to be observed and recorded.

Halo effect: The tendency to let the rating of one characteristic of a person to be influenced by another characteristic or by one's general impression of that person.

Behavioral description: A statement or series of statements of a behavior for groups or individuals.

Behavioral scale: The classification of a teaching event into behavioral descriptors that can be observed, as applicable to specific district goals.

Objectivity: The degree to which the instrument is

free of influence or distortion by beliefs or biases of the person using it.

Instrument: The instrument for the observation model for evaluating instructional performance consisting of eight to twelve behavior description scales which define the instructional activities of the teacher and provide the basis for assessing the classroom behaviors of teachers.

Scalogram: For the purpose of this project, each behavior description scale was composed of three items stated in behavioral terms. The three items range from most desirable to least desirable teaching behavior.

Review of the Literature and Research

A review of the literature and research was conducted to identify highly reliable performance evaluation methods. In addition, an attempt was made to identify performance evaluation methods that obtain objective analysis of instruction and have the most potential for local school district use. The review of the literature is presented in chapter two.

Methodology of the Study

Field Test Sites

There were two urban and one rural school districts that participated in field testing the model. The three school districts developed performance evaluation instruments

according to the model's specifications. The methodology is presented in chapter three.

Treatment of the Data

The data collected from the field test site evaluation of instructional personnel was treated for relationships among the independent raters. Teams of three observers in each field test site conducted a minimum of three performance evaluations in actual classroom settings to test the results for reliability. A Pearson Product Moment coefficient of correlation for the data from the observations was used to determine the reliability level. A percentage of agreement between raters for each instructional staff observed was also applied. These data provided information about the strength of the relationship among raters.

Method of Presentation of Results

The information and data collected from the field test sites is described in narrative form and visually displayed through the use of tables in chapter four. Additionally, samples of each district's instrument and scales developed and tailored to their own district goals, are incorporated into the appendices.

The data are summarized, conclusions drawn and recommendations presented in chapter five.

Importance of the Study

This study offers school districts a model for developing a reliable evaluation instrument. An evaluation model that provides for objective and reliable data of instructional personnel will serve to encourage school districts to be more optimistic about implementing evaluation procedures. For a school district to use the model, local goals or objectives will need to be available or established. The project should have meaning to school district administrators, instructional personnel, and boards of education, the education community, and university educational administration training programs.

CHAPTER TWO

Review of the Literature

The focus on the review of literature was to identify the current methods of evaluating teacher performance, including rating systems, student achievement systems, personality traits and teaching process behaviors. Current legislation effecting teacher evaluation was reviewed. Observational methods were looked at regarding positive and negative factors of reliability, validity, training and involvement of instructional personnel.

A wide variety of methods for developing performance evaluation instruments are available. Some compare the teacher's performance against other teachers' performances, and others compare the teacher's performance against established standards of performance. One source of resistance to performance evaluation arises from a lack of awareness about how evaluation fits into an overall model for the effective management of people.[1]

A recent movement by researchers has been the

---

[1] L. L. Cummings and Donald P. Schwab, Performance in Organizations (Glenview, Illinois: Scott, Foresman, 1973), p. 6.

development of systems to ensure that observation procedures are described objectively.[1] The instruments and methods in use for actually describing teaching behaviors are fairly small in number and most are more suitable for research in social psychology than performance evaluation.[2] Gallagher, Nuthall and Rosenshine point out that "what most evaluators need is some easy-to-administer, easy-to-understand, not very controversial valid method of describing teaching."[3]

The school districts of this nation are experiencing strong demands for accountability. Schools are being required to prove they are using the money they have wisely. An important component of this accountability movement is personnel evaluation. Lamb and Swick report that observation of teaching behaviors has emerged as an evaluation technique for addressing the accountability demands.[4]

Evaluation of instructional performance is a difficult task for school administrators. Board of education policies, community insistence and state legislation requires that

---

[1]Walberg, pp. 1-9.

[2]James J. Gallagher, Graham A. Nuthall, and Barak Rosenshine, Classroom Observation (Chicago: Rand McNally, 1970), pp. 4-6.

[3]Ibid., p. 5.

[4]Morris L. Lamb and Kevin Swick, "A Historical Overview of Classroom Teacher Observation," Education Forum, 39 (January 1975), 239-47.

instructional personnel performance be evaluated. Bishop maintains that the research on observation instruments has provided impressive data and the potential for powerful tools, but that the data have not been adequately considered.[1] Many performance evaluation tools are still in use that provide results that are virtually worthless. They may indeed reflect more of the subjective view of the evaluator than a true measure of the individual being evaluated.

Administrative personnel who evaluate instructional personnel, will no longer be able to take shelter in outdated evaluation instruments and procedures.[2] The decision for school administrators is not whether to evaluate performance, but rather how to do so and what methods to use.

Performance evaluation of instructional personnel has challenged researchers and practitioners, stimulating a half-century of research, but few are satisfied with the state of the art.[3]

---

[1]Leslee J. Bishop, "Systems for Observing In-School Operations," in Observation Methods in the Classroom, eds. Charles Beagle and Richard Brandt (Washington, D.C.: Association for Supervision and Curriculum Development, 1973), p. 9.

[2]Kenneth Dunn and Rita Dunn, Administrators Guide to New Programs for Faculty Management and Evaluation (West Nyack, New York: Parker Publishing, 1977), pp. 205-19.

[3]Stephen J. Knezevich, "Designing Performance Appraisal Systems," New Directions for Education, 1 (Fall 1973), 37-50.

Evaluating Teacher Performance

Teacher evaluations have traditionally focused on different techniques. There have been three commonly used techniques: (1) efficiency ratings, (2) pupil growth, and (3) pre-service criteria.[1]

Early researchers focused their attention on the broader concept of classroom climate rather than characteristics of instructional personnel. Their studies proved interesting, but not manageable in instructional evaluation. Descriptors such as dominative/integrative and learner-centered/teacher-centered proved difficult to define or observe. There are those who support the earlier contentions of Bellack et al. that studying the effectiveness of instructional personnel in terms of teacher variables is futile.[2]

More recent studies, e.g., Rowe, indicate that specific and well defined teacher variables can be generated.[3] The first step in the development of any performance evaluation

---

[1]Arvil S. Barr, Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness: A Summary of Investigations (Madison, Wisconsin: Dembar Publications, 1961), pp. 23-47.

[2]Arno A. Bellack et al., The Language of the Classroom (New York: Teachers College Press, Columbia University, 1966), pp. 41-86.

[3]Mary B. Rowe, "Wait, Time and Rewards as Instructional Variables," Journal of Research in Science Teaching, 11, No. 2 (1974), 81-94.

system should involve a careful analysis of the expectations of the job.[1] The final step involves the actual development of the instrument.

Rating Systems

In studies reviewed by Rosenshine and Furst, the most consistent results and the highest correlations were obtained with rating systems.[2] Their review contained the conclusion that the research to date on direct observations in natural settings is so varied that one method or format could not be presented as superior to another. They clarify that it would be a misrepresentation of their review to consider these results as evidence in favor of rating systems or against category systems. (When an event is recorded each time it occurs it is labeled a category system, and when observers estimate frequency of events only once at the end of a session, it is referred to as a rating system.)

It is reported in the literature that in a majority of school systems, classroom observation of teachers by principals using a checklist or rating scale is the standard method of evaluating performance. Rosenshine and Furst conclude that this practice will continue.[3] Morrison and McIntyre summarize the case against rating scales thusly:

---

[1]Cummings and Schwab, pp. 55-69.

[2]Rosenshine and Furst, pp. 129-83.

[3]Ibid.

> Despite their popularity several objections can be raised against rating scales. One of the more serious limitations when used for assessing the classroom behavior of teachers is that an extensive amount of information about what has gone on has to be reduced to subjective and impressionistic endorsements on a few scales. Since they are heavily dependent upon the subjective impressions formed by the individual rater, their reliability from one occasion of rating to another by the same rater, or between two or more raters on the same occasion, is highly variable. Also, when the rater is presented with several supposedly distinct characteristics to assess he may in fact be unable to distinguish between them, leading to a tendency to rate an individual as "high," "average" or "low" on most of them. Finally the information available to the rater can vary very much from one characteristic to another and from one individual to another.[1]

The common approach of having principals determine teacher performance from rating scales is unacceptable in view of their being too susceptible to bias.[2]

Behavioral rating scales have the potential to overcome many of the problems found in other methods.[3] The developmental process generally involves both the employees and the administrator. Once the categories for performance are identified and agreed upon they are included into the measuring instrument for field testing. Categories for

---

[1]Arnold Morrison and Donald McIntyre, Teachers and Teaching (Baltimore: Penguin Books, 1969), p. 22.

[2]Stephen Klein and Marvin C. Alkin, "Evaluating Teachers for Outcome Accountability," in The Appraisal of Teaching: Concepts and Process, ed. Gary D. Borich (Reading, Massachusetts: Addison-Wesley, 1977), pp. 231-33.

[3]Ibid.

performance evaluation are behavioral activities and must be capable of being observed and with terms having the same meaning to independent raters.

Rating scales tend to be more haphazard than systematic observations. With ratings, the rater is generally required to make some sort of evaluative judgment as to whether a given aspect of the teachers performance is good or bad. An individual using a systematic observation approach is only required to record whether a specific behavior occurred.

A most promising aspect of the behavioral rating approach pertains to its potential value for employee development through feedback, since fairly specific behavior is pinpointed in the observation process.[1]

Medley and Hill contend that "further study in the scientific study of teaching depend on the development of practical objective procedures for measuring teaching behaviors."[2]

Student Achievement Systems

Given the present state of the art of performance evaluation, researchers Herbert and Smith question the appropriateness of using student outcomes to measure teacher

---

[1]Cummings and Schwab, p. 95.

[2]Donald M. Medley and Russell A. Hill, Measurement Properties of Observation Schedules and Record (ERIC ED 185 089), p. 122.

performance.[1] The research on teacher effectiveness has had little success in identifying the characteristics of effective teachers associated with success in student achievement gains. Brophy comments that "the positive findings which have appeared are weak ones."[2]

Greer says he is tired of the claim that there is no valid method of identifying the most qualified teachers.[3] As an alternative he suggests that "serious" students be asked to identify the best teachers. But on what grounds such judgments are to be made are not provided by Greer. One would have to question the use of only "serious" students rather than all students. Regardless of the concerns of his procedure, his statement does reflect a contemporary plea for improving the public schools performance evaluation methods. Borich, Malitz and Kugle report that "researchers have used observation instruments for the greater part of a decade to investigate the relationships between teacher behavior and student outcome with few consistent

---

[1]John Herbert, "A Research Base for Accreditation of Teacher Preparation Programs," Accreditation and Research Problems, eds. John L. Burdin and Margaret T. Reagan (ERIC ED 050 021), pp. 3-24, and B. Othanel Smith, Certification of Educational Personnel (ERIC ED 055 975), pp. 5-17.

[2]Jere E. Brophy, Stability in Teacher Effectiveness (Austin, Texas: The Research and Development Center for Teacher Education, University of Texas, 1972), p. 2.

[3]Peter R. Greer, "Another Simple Truth," Education Week, 1 (June 1982), 20.

findings."[1]

During the 1970's much emphasis was placed upon criterion referenced tests and student achievement outcomes which may account for the limited amount of literature on observation of instructional behaviors during this same period. There was only one reference in the Review of Educational Research from 1972 to 1979 regarding observational measures. Many authors maintain that factors other than the influence of the teacher contribute significantly to changes in pupil behavior and, thus, it is not possible to evaluate the work of a teacher solely in terms of the achievement of the pupils. Teachers and administrators are well aware of procedures for examining pupils to assess their learning. They have generally been unwilling, however, to use the results of these assessments as a basis for evaluating instructional personnel. There are several reasons for this, some of which have been previously presented.

These reasons can probably be summarized on the basis of: (1) there being too many factors which affect the amount of a student's learning outside the instructional setting, and (2) the tests selected may not reflect the entire range of goals for which the district's educational program is concerned. Klien and Alkin quip "therefore, the use of most

---

[1]Gary D. Borich, David Molitz and Cherry L. Kugle, "Convergent and Discriminant Validity of Five Classroom Observation Systems: Testing a Model," Journal of Educational Psychology, 70 (April 1978), 119.

nationally normed standardized tests to assess a given teacher's performance would be analogous to using a bathroom scale to determine how many stamps to put on a letter."[1]

Glass bases his opposition to the use of pupil outcome on statistical grounds. After examining twenty-one studies, he points out that the findings support arguing against both standardized achievement tests and teacher performance tests for determining the quality of instructional performance.[2]

Personality Traits

To emphasize the influence that personality traits continue to have on instructional evaluation systems, some authors still support their inclusion as part of an effective performance evaluation system. Marks states "it was the only system of evaluation of teacher effectiveness for many decades, and is still an important force with which to reckon."[3] Herman provides a more positive use of personality traits by stating "local planners need to decide what personality characteristics and what quantity of work produced

---

[1]Klein and Alkin, p. 232.

[2]Gene V. Glass, "A Review of Three Methods of Determining Teacher Effectiveness," in The Appraisal of Teaching Concepts and Process, ed. Gary D. Borich (Reading, Massachusetts: Addison-Wesley, 1977), pp. 224-341.

[3]Merle B. Marks, "Effective Teacher Evaluation," National Association of Secondary School Principals Bulletin, 60 (September 1976), 6.

are considered acceptable."[1]

Teaching Process Behaviors

In the area of process behaviors, significant research advances seem to have been made in recent years. The process behaviors relate directly to teaching activities and include variables of teacher behavior and teaching activities in the classroom. Among the limitations that have been encountered in the use of process behaviors have been the difficulties of recording behaviors and categorizing them objectively and reliably. The studies of Medley, Mitzel, Ober, Popham, Borg, Hyman, and others indicate that comprehensive and significant research in the area of teacher behavior has been initiated in the past decade.

The California Teachers Association illustrate observable evidence of teacher performance as follows:

> Emphasis is placed upon results to be accomplished, rather than on "how to do it." We are concerned with what the teacher must be able to do, not how he is to do it. In adapting to specific situations, an expert teacher may use any one of several techniques. Professional expertness is not defined meaningfully through job analysis and compilation of procedures used by teachers known to be successful. Professional practice is based upon expert diagnosis and choice among techniques, adaptation of known techniques, or development of new techniques.

---

[1] Jerry J. Herman, "Developing a Staff Evaluation Program," National Association of Secondary School Principals Bulletin, 60 (September 1976), 9.

> Blind imitation of a model, no matter how "expert" is fatal to professional practice.[1]

These definitions are not specific or universal. Additional refinements and adaptation to local district goals would be necessary. Agreement on goals and procedures should precede any development of an instrument for evaluating instructional performance.

Legislation

The New Jersey Education Association has alerted the teachers they represent to the state's mandated evaluation policy through an article in their October 1979 journal. The article alerts the teachers to be knowledgeable of what will be evaluated, to be prepared, and to provide pointers for the post evaluation conference.

With high inference evaluation instruments, sooner or later a teacher receiving a low rating, and as a result terminated, will go to court and question the inferences made. When this occurs the district will have to rely heavily upon the expertise of the rater. Any charge of bias or unreliability would be difficult to refute. Specific requirements for the evaluation of instructional personnel exists in fewer than half of the states.[2] State laws concerning

---

[1] Six Areas of Teacher Competence (Burlingame, California: California Teachers Association, 1964), p. 17.

[2] "Teacher Evaluation," A Legal Memorandum (Reston, Virginia: National Association of Secondary School Principals, 1978), p. 4.

teacher evaluation are generally written in conjunction with other legal issues. They are usually interlocked within the context of teacher tenure, collective bargaining or teacher certification. The degrees of specificity in the states vary widely.[1] In the majority of the states no requirement is made for uniform, standardized or objective measures. In the states where local discretion is permitted, districts would be wise to establish clear standards and procedures for evaluation of instructional personnel. Their failure to do so may lead to unnecessary and time consuming legal reviews.[2]

More and more states are enacting legislation to establish requirements for evaluating public school personnel. New Jersey, Ohio, North Carolina and Nebraska are four of the recent ones enacting legislation in 1982.

The North Carolina plan was approved by that state's board of education to begin in the fall of the 1982-1983 school term. Two years ago, the North Carolina general assembly directed the state board of education to develop a statewide plan. State officials describe the system as one in which the state will leave most of the evaluation process and use of its results up to local districts. The professional skills for which teachers will be evaluated

---

[1] A Legal Memorandum, pp. 2-6.

[2] Ibid., p. 4.

include planning, giving and overseeing instruction, identifying student strengths and weaknesses, and working well with colleagues.[1] Guidelines are not provided, however, on what constitutes satisfactory performance. This will be left up to the discretion of each local school district.

Nebraska's Legislative Bill 259 primarily spells out due process procedures for instructional personnel. Contained within the bill are specific regulations governing the evaluation of instructional personnel. Legislative Bill 259 specifically states

> All probationary certified employees employed by Class I, II, III and VI school districts shall, during each year of probationary employment, be evaluated at least once each semester in accordance with the procedures outlined.[2]

The procedures prescribe that evaluation be based upon observed classroom instruction for at least an entire instructional period. Probationary certified employees are defined as those who have served under a contract with the school district for less than three years. An informal discussion with several Nebraska school administrators at regional meetings reflects that L.B. 259 will influence the school districts to evaluate all professional employees' performance in the event change of status is warranted.

---

[1]Alex Heard, "N. C. To Begin Statewide Evaluation of Teachers, Principals," Education Week, 1 (August 25, 1982), 6.

[2]Nebraska, Legislative Bill 259 (1982), pp. 1-10.

The motivation for state governments to enact legislation requiring the appraisal of instructional personnel has largely stemmed from the accountability and community concern movement more than for the purpose of improving instruction. As might be predicted, state generated legislation has been received warmly by the citizens, received with caution by school administrators, and looked upon by teachers with skepticism.

Evaluation of instructional performance will be used more in the decade to come to document administrative decisions for legal protection and for development and training.[1]

According to Hyman, the various new laws and court decisions require principals and other supervisors to be more judicious, more careful, more sure of their data, more precise and more helpful.[2] This is significant because many school districts realize their procedures for gathering data, as well as the kind of data they have used, will not stand up in court.[3] This suggests that school districts must improve their performance evaluation procedures and

---

[1]Ann Morrison and Mary Ellen Krantz, "The Shape of Performance Appraisal in the Coming Decade," Personnel, 58 (July-August 1981), 12-22.

[2]Ronald T. Hyman, School Administrators Handbook of Teacher Supervision and Evaluation Methods (Englewood Cliffs, New Jersey: Prentice-Hall, 19765), pp. 7-9.

[3]Ibid.

carefully observe the instructional personnel.

Current case law regarding teacher evaluation is difficult to report because of the variety of state and local provisions. The scarcity of cases may be partially due to the recency of state statutes for teacher evaluation. One conclusion that does appear relevant is from the Legal Memorandum, "Courts tend to strictly apply the procedural requirements of teacher evaluation laws. Principals who fail as evaluators may, themselves, fail as evaluatees as a result."[1]

Caldwell stated in her article on the 1981 Seminar of National Organizations on Legal Problems of Education (NOLPE) that school boards most often blunder on procedure, and that procedure is the first question judges consider in law suits over personnel actions.[2] Administrators are singled out as shirking their responsibilities in failing to implement these procedures. Three specific suggestions were offered by NOLPE to keep schools out of court: (1) a written district wide policy on employee evaluation, (2) a simple clear procedure for evaluation, and (3) multiple observations by more than one person.[3]

---

[1] A Legal Memorandum, p. 8.

[2] Peggy Caldwell, "Teacher-Evaluation Methods Called Inadequate," Education Week, 1 (November 1981), 4.

[3] Ibid.

Observational Methods

Factors Against

Besides the difficult job of evaluating instructional personnel in general, there are particular problems in many performance evaluation procedures that can reduce the effectiveness of evaluation. The individuals doing the evaluating or rating may make certain errors. They may rate the individuals too high which is referred to as leniency error, or they may not discriminate sufficiently among staff and give everyone a similar rating. Ratings may be based upon a single key trait or aspect of the job rather than an evaluation of all the important goals separately e.g., an employee who is often tardy may likely be rated low overall, even if performance is good in other areas. Many evaluators also do not know what is required of them or what the important criteria or goals are of the job.

According to Dunn and Dunn, "administrators have been evaluating teachers for many years in an effort to isolate those characteristics that produce effective instruction."[1] They identified two weaknesses of performance evaluation attempts: (1) incorrectly identifying common characteristics of teacher performance, and (2) failure to objectively interpret what was observed.

Too many performance evaluation procedures become

---

[1]Dunn and Dunn, p. 150.

entangled with human relations and concentration on personality, such as rapport with others, appropriateness of dress, or cooperation with the administration.[1]

Common types of performance evaluation errors are: (1) prejudice, (2) halo effect, (3) leniency, (4) central tendency, (5) tenure effect, (6) judgment of persons over short period of time, and (7) level of prior performance.[2]

Kult points out that

> rating systems and criteria dependent upon arbitrarily interpreted categories as enthusiasm, pleasant appearance and loyalty to the school promise little if any helpful teacher measurement, and can even do potentially more harm than good.[3]

Two critical problems persistently impede the appraisal of teaching according to Borich: "(1) the collection of data that are reliable, and (2) the presentation of these data to the teacher in an accurate and comprehensive form."[4]

To gain the board of education and staff support for performance evaluation of instructional personnel,

---

[1]James W. Popham, Educational Evaluation (Englewood Cliffs, New Jersey: Prentice-Hall, 1975), pp. 284-90.

[2]Walter B. Roettger, Performance Appraisal Skills for Managers and Supervisors (Des Moines, Iowa: Institute of Public Affairs and Administration, Drake University, 1981), p. 30.

[3]Lawrence C. Kult, "Improving Teacher Evaluation by Principals," The Clearinghouse, 52 (September 1978), 18.

[4]Gary D. Borich, ed., The Appraisal of Teaching: Concepts and Process (Reading, Massachusetts: Addison-Wesley, 1977), p. 45.

Egglebrek suggests that it must be approached in a positive way and through the development of goals and objectives.[1] He also emphasized that evaluation will be much better if used for improvement of instruction.

Prejudice and halo effect are particularly susceptible to both how the person conducting the evaluation feels about the person being evaluated and past performance of that person. Once teachers have been regarded as very good or very poor, the tendency is to continue to view them the same way. Rating scales employing rankings of very good, average, and very poor are often identified with lenient evaluators or those who tend to rate most people as average.

Medley and Mitzel contend that observation of classroom behavior is seldom included in research studies because:

1. Observations are expensive.
2. They constitute invasion of privacy.
3. They are disturbing and may cause atypical behavior.
4. The methodology has not increased knowledge about teaching and learning.[2]

It should be noted that these researchers' interest in observation was research oriented while the major interest of classroom observation of behaviors of instructional

---

[1]Dave Egglebrek, Evaluation of School Instructional Programs: How Do You Do It?, cassette (Washington, D.C.: Educational Resource Information Center, April, 1981).

[2]Donald M. Medley and Harold E. Mitzel, "Measuring Classroom Behavior by Systematic Observation," in Handbook of Research on Teaching, ed. N. L. Gage (Chicago: Rand McNally, 1963), pp. 247-328.

personnel is for evaluation, improvement of instruction and professional development plans.

Factors For

Performance or behavioral based rating scales are a fairly new development that could greatly increase the effectiveness of appraisal programs.[1] Highly reliable, low inference observation scales can be used to collect data from which patterns of teacher behavior can be identified. The most promising application of behavior observation instruments is in the area of feedback to the teacher.[2] Because of the specificity of observed classroom activities, teachers can relate this to areas of interaction with their students. Feedback to teachers based on behavioral observation appears to be an improvement over statements like "you seem to relate to the students" or "you do not seem to be able to motivate students."[3] As a feedback mechanism, actual observed, objective data provides teachers something concrete upon which they can build.

According to Roettger, the elements of a good

---

[1]John D. McMillan and Hoyt W. Doyle, "Performance Appraisal: Match the Tool to the Task," Personnel, 57 (July-August 1980), 12.

[2]James A. Shymansky, "Assessing Teacher Performance in the Classroom: Pattern Analysis Applied to Interaction Data," Studies in Educational Evaluation, 4 (Summer 1978), 99-106.

[3]Ibid., pp. 99-106.

evaluation system include:

1. objectivity
2. reliability
3. validity
4. high discriminability
5. standardization in form and administration
6. training for raters and ratees
7. practical and cost effective
8. mechanisms for appeal
9. standards and expectations in writing
10. standards and expectations mutually agreed upon.[1]

Classroom observation systems are one type of evaluation of instructional performance that will overcome most if not all of these common measurement errors. Kugle claims that

> classroom observation systems provide a direct, quantitative account of classroom activity, supplying a more objective means of analyzing teacher behavior than other process measures, such as teacher, peer or administrator ratings.[2]

Use of classroom observation to identify effective teaching behaviors has proved to be a promising technique, at least in situations where the same content is taught to different pupils and where classroom observations focus upon instructional activities.[3]

Among the advantages cited for systematic observations is that of Popham. He states that "their advantage is with

---

[1]Roettger, p. 18.

[2]C. L. Kugle, Data Collection Procedures for the Evaluation of Teaching Program, Phase III (ERIC ED 170 340), p. 2.

[3]Borich, The Appraisal of Teaching, pp. 8-14.

the reliability with which such observations can be made."[1]

Observational systems do not produce evaluative judgments, but rather, serve as tools for obtaining data that can be used for determining what actually happens in the classroom.

Ober, Bentley and Miller state that "accurate and objective feedback can be possible through techniques of systematic observation."[2]

Systematic observation provides a means for focusing on the variables interacting in an instructional situation. It looks specifically at the process of facilitating learning. Observational instruments are also reported to be more objective and therefore are more likely to be acceptable by the teachers than other methods such as opinion or rating scales.

Ober et al. introduced systematic observation techniques to hundreds of teachers, principals and other professionals and confirmed that time, money, and skill are in a relative sense, very inexpensive.[3]

Soar's research suggests that a manageable amount of different kinds of teacher behaviors is best for a particular

---

[1]Popham, p. 289.

[2]Richard L. Ober, Ernest L. Bentley and Edith Miller, Systematic Observation in Teaching (Englewood Cliffs, New Jersey: Prentice-Hall, 1971), p. 89.

[3]Ober, Bentley and Miller, pp. 57-68.

goal.[1] He takes exception to several of the traditional teacher rating techniques, such as administrative ratings and student gain methods, and suggests the measurement of teacher process. He indicates that systematic observations of teacher behaviors based upon objective instruments are appropriate.

Both Scriven and Popham agree that the setting of priorities and goals by individuals involved in the performance evaluation process is the most useful way to determine the behaviors for which the teacher will be held accountable.[2] The research literature changed in the 1960's to begin providing measures of teacher behavior for clarifying the nature of teacher effectiveness, although this procedure is very complex. The more promising findings have come from the use of systematic observations rather than rating procedures. The use of systematic observation could meet the requirements of an evaluation model for teacher competencies to be determined from clearly stated expectations or goals and be made public in advance.[3]

---

[1]Robert J. Soar, "An Integration of Findings from Four Studies of Teacher Effectiveness," in The Appraisal of Teaching: Concepts and Process, ed. Borich, pp. 95-103.

[2]Michael Scriven, "The Evaluation of Teachers and Teaching," in The Appraisal of Teaching: Concepts and Process, ed. Borich, p. 96; W. James Popham, Educational Evaluation (Englewood Cliffs, New Jersey: Prentice-Hall, 1975), pp. 261-77.

[3]Borich, pp. 166-73.

Involvement of Instructional Personnel

The morale of instructional personnel seems to deteriorate and hostility often develops among them when they are not involved in the development of performance evaluation procedures. Because most evaluation instruments reflect the philosophy of the persons who design them, it would be appropriate for both administrators and teachers to determine the contents for the evaluation instrument.

Mutually agreed upon objectives will provide the basis for an acceptable and meaningful performance evaluation instrument by both teachers and administrators. Robinson and Lee think that agreement on the teaching behaviors has a direct and positive effect on the task of teacher evaluation.[1]

One of the major goals in developing an instructional evaluation model is to involve teacher input. Gootas and Rutherford advocate involving teachers in the development of the evaluation procedures, by permitting those teachers input as to what and how they will be evaluated.[2]

If employees are involved in the performance evaluation process from the beginning and are active participants in the outcome the chances are much better for their support.

---

[1]John J. Robinson and John H. Lee, Jr., "Evaluation: Can We Agree?", National Association of Secondary School Principals Bulletin, 62 (December 1978), 15-20.

[2]Harry Gootas and Jerry Rutherford, Professional Evaluation for Teachers--Policy to Practice (ERIC ED 188 324), pp. 9-10.

When employees are given an opportunity to become more proactive in the appraisal process, they require far less attention from the administration and can make a greater contribution to the overall effort of the district's goals.[1]

Training of Observers

Highly skilled and trained observers are an essential component of the evaluation of instructional personnel. Observers must record specific behaviors as they occur, interactions in a classroom cannot be rerun to check behaviors.

Ward stated, "whenever an assessment program demands high accuracy in recording teacher performance, specially trained individuals will be needed."[2] Both accuracy and judgment in classifying the behaviors of instructional personnel correctly are necessary qualities of a good observer. Ober, Bentley and Miller express that "unless recorded behaviors are actual, observed behaviors, a system's usefulness is limited: The greater the disparity between observed and recorded behaviors, the less useful the system."[3]

---

[1]Beverly L. Kaye and Shelly Krantz, "Preparing Employees: The Missing Link on Performance Appraisal Training," Personnel, 59 (May-June 1982), 23-29.

[2]Beatrice A. Ward, Assessment of Teacher Performance: What is Involved? What is the Cost? (ERIC ED 177 150), p. 15.

[3]Ober, Bentley, and Miller, p. 79.

Robinson noted that most evaluators had no training in observational techniques, and did little or no preparation before observing a teacher.[1] Training observers for collecting data on teacher behaviors is more likely to be valid than data collected by untrained observers. This is particularly so when the observers have been trained to identify specific and predetermined behaviors of the instructional personnel. If observers are not properly trained they are much more likely to gather data based on their own biased perceptions and subjective interpretations. Individuals are also susceptible to carrying over a general impression they have regarding one behavior to all other behaviors they are observing. Fewer inferences of evaluated behaviors will result if observers are properly trained. Van Dalen claims "the more training an observer receives the less variation will occur between raters and the higher the reliability of the evaluations, especially when evaluating marginal and difficult behaviors."[2] When the defect of rater training is remedied, ratings reach very high levels of consistency and are stable over long periods.

---

[1]John Robinson, "The Observation Report--A Help or A Nuisance?" National Association of Secondary School Principals Bulletin, 62 (December 1978), 22-26.

[2]Van Dalen, p. 348.

Reliability of Observational Methods

Sirotnik notes that for observation techniques of performance evaluation, the circumstances due to different observers or different observational occasions are particularly important for the existence of reliability.[1] With regard to observation of instructional personnel, the reliability coefficient of correlation between different observers on the same occasion has high utility. The inter-observer reliability is particularly useful as an index of the objectivity of a model and a desirable ingredient for evaluating a model. If a system is to be used reliably by raters, scale descriptions must be clearly operationalized for the observers. This of course, is the purpose of training and is substantiated in the section which discusses training of observers. It should be noted that it is not always possible to include in a definition manual all the information necessary to record a particular category reliably.[2]

Reliability appears to be a component missing in several observation methodologies. It is essential for performance evaluations to be reliable if any generalization is to occur. Poorly designed models for evaluating instructional

---

[1]Kenneth A. Sirotnik, "An Inter-Observer Reliability Study of the SRI Observation System as Modified for Use in a Study of Schooling," in A Study of Schooling (Los Angeles, California: University of California, Technical Report No. 27, 1981), pp. 119-28.

[2]Ibid.

performance will affect the reliability. From Frick and Semmel's viewpoint, systems with high inference items cause problems in observer agreement from the very beginning when the observers are being trained.[1]

The word reliability when applied to observational data is usually interpreted as referring to observer agreement. Medley and Mitzel regard reliability in measurement as occurring "when the purpose of observations is for obtaining an accurate description of what happened in the classroom at a particular time."[2]

Validity of Observational Methods

Performance evaluation systems must be concerned with the aspect of validity of their instruments. Ober, Miller and Bentley consider the inter-rater agreement between observers as an index of validity.[3] For an observational scale to be valid for measuring behavior, it must provide an accurate record of behaviors which actually occurred, scored in such a way that the scores are reliable.[4]

---

[1]Ted Frick and Melvyn I. Semmel, "Observer Agreement and Reliabilities of Classroom Observational Measures," Review of Educational Research, 48 (Winter 1978), 157-84.

[2]Medley and Hill, p. 7.

[3]Ober, Bentley, and Miller, p. 82.

[4]Ibid.

Content validity can be shown if the evaluation instrument contains an adequate and representative sample of the duties and responsibilities necessary for performance on the job. Items must be chosen on basis of the degree to which they cover the range of important behaviors of the job and the extent to which they discriminate between good and poor performance.

If one is trained to record data and to recognize specific behaviors and does not interpret or draw inferences, the procedure usually produces very reliable and valid results.[1]

Evaluation based on standards of teacher behaviors are validated primarily in terms of the adequacy with which they represent the standards. Therefore, content validity approaches would be more suited to such evaluation procedures. Validity can be improved through increasing the number of occasions the instrument is tested.[2] This is generally done by finding an optimal number of occasions and observers to increase the reliability and thus increasing the validity.

Discussion

A review of the literature indicates that the most comprehensive performance evaluation system is that of

---

[1]Borg, pp. 243-52.

[2]Popham, pp. 153-60.

employing systematic observation procedures. (1) Observers require training to become skilled at recording and scoring instructional performance. (2) Specific behaviors must be identified and recorded at the moment they occur. Instructional activities within a classroom cannot be played back for the observer to check her/his recordings. (3) Classroom observations appear to be most useful for assessing instructional performance. (4) The particular behaviors to be observed should be in a natural setting for the teacher and performance can then be evaluated as soon as it occurs.

The several evaluation methods that have been reviewed in the literature contain various techniques in format, purpose and procedure. Some of the current approaches in evaluating instructional personnel include rating scales, student rating of teachers, student achievement, checklists, comparison techniques and behavioral observation methods.

Given the variation that exists in evaluation instruments, it would seem unwise to limit a model to a single system. The optimal strategy would be to use classifications from a variety of conceptual or theoretical base instruments. Observational instruments for performance evaluation as identified in the literature, offer the potential for evaluating instructional behaviors at different levels of specificity. If such instruments are viewed as tools for measurement, then those most useful for evaluating classroom performance can be modified.

The performance evaluation of instructional personnel, for the purpose of instructional improvement and professional development, requires that teachers be actively involved in the development of the evaluation process. Their involvement will lend credibility to the evaluation program and it will more likely be viewed by them as having value. If teachers are not involved they will most likely lack the commitment and ownership in the evaluation program to help make it a success. Instructional personnel are in a unique position to circumvent the usefulness of the evaluation program. Evaluation programs must be developed which are flexible and yet structured. They should follow a consistent format or model in order to meet the needs of the community, the administrators and the instructional personnel.

The following six criteria provide a thrust for developing a viable and improved model for systematic observation of instructional personnel:

1. The procedure should be identified with the community and district goals.
2. The procedure should be identified with the instructional setting.
3. The specific teacher behaviors and how they will be evaluated should be clearly stated.
4. Improvement of instructional and professional development should be the primary objective.
5. The evaluation procedure should be cooperatively

developed by the instructional and administrative staff.

6. Evaluation and review of the system should be an ongoing procedure.

Instructional performance evaluation is or should be of high interest for all members of the school community. Since there does not appear to be any definitive research that provides direction to a perfect evaluation system, the task of developing, modifying or selecting a plan will be the responsibility of the local school district.

If the information obtained from evaluating instructional personnel is to have any use, there must be some specific procedure for interpreting the data that is gathered. Without this procedure the data is only descriptive with little value to either the district or the teachers.

Of prime significance in an instructional setting is the nature of the goals and objectives toward which the instructional activities are directed. Goals vary from one school district to another and from one classroom to another. As goals vary, the importance of the various instructional behaviors will also vary. Thus the first step in developing an instructional evaluation model is to ensure the availability of district goals. If not available, then district goals must be developed.

Borich's caution that evaluation is an area that still

needs work, seems appropriate.[1] In comparison to the other alternatives reviewed, systematic observation methods seem to have the most potential.

Teachers have had serious misgivings about evaluation systems and have a tendency to distrust the process. This is probably to be expected in view of the past use of subjective, judgmental, high-inference and trait related measures.

An important factor not evident in the various readings of the literature is an observational system that lends itself to systematic implementation at a local school district level.

Systematic observation of instructional activities with the use of an instrument for guiding the observer, appears to be quite rare at present. Several instruments have been developed, but are weak in reliability.

In the literature review it was noted that research studies ranged from measuring single criterion of behaviors to broadly defined teacher behaviors, providing little opportunity for objective observations. Little information was obtained regarding the replicability and uniformity of instructional evaluation instruments.

---

[1]Borich, pp. 171-73.

Summary

This chapter reviews the various methods and procedures for establishing programs for evaluating instructional personnel. A wide range of choices are evident, but the usefulness of the data collected varies significantly among the methods. The fact that well conceived instructional evaluation programs may exist, does not mean that the school can automatically apply the procedures to their district. Other schools may not be aware that conceptual programs exist.

The tentativeness of several conclusions arrived at in the research on teacher evaluation regarding their applicability to specific school districts might suggest the need for caution. The practical problem of evaluating teachers is a pressing issue. It is essential that research findings be used even though they are tentative.[1]

Since evaluation systems for instructional performance are a reality, it would be better to develop good ones rather than have poor ones. The literature provided data on what a system should be like. It should not require subjective judgments by principals or other observers about the teacher's performance. A good model will emphasize objective and observable teacher behaviors for determining performance. It will also be concerned with the goals of the

---

[1]Peter Coleman, "The Improvement of Aggregate Teaching Effectiveness in a School District," in The Appraisal of Teaching: Concepts and Process, ed. Borich, pp. 217-29.

education community for which it is to be used. If the instructional activities do not relate to the goals of the education community, the quality of instructional performance is of little value.

From the review of the literature, it is difficult to draw specific conclusions or directions observation of instructional personnel will move in the future. With the current concerns revolving around community and teacher involvement, accountability, and performance, the use of observation of instructional behaviors is moving toward an educational priority.

CHAPTER THREE

Methodology

This project was to develop a model for evaluation of instructional personnel. It calls for greater involvement of community, administrators and teachers and increased communication and cooperation among the administrative and instructional personnel.

Development of the Model

The methodology applied in the development of this model was an alternate approach to the usual practice of school districts using rating scales as discussed in the review of literature. A conceptual approach was developed based on past experiences, graduate studies, and review of the literature. The application of the model required taking instructional process behaviors of teachers and then observing these to identify teaching performance. The topic of performance evaluation is abundant in the literature, but a procedure was needed for putting it all together into an operational evaluation model, especially one that would be manageable, practical and district specific.

The model provided the framework for a specific series of actions to be followed by the local school districts. When the districts followed the series of actions, they had

a completed performance evaluation instrument. The model established the type of content for the instrument and the districts developed the specific content. The district specific content had to include district goals to be used for individual scale topics in the instrument. These topics were then defined and a behavior scale was written for each. Teaching behaviors were written for each behavior scale completing the series of actions.

The emphasis of the model was to allow school districts to integrate their educational goals into the best elements of performance evaluation. The project was aimed at individuals who are responsible for the performance evaluation of instructional personnel. These practitioners had to consider the priorities of the community, administrators and instructional staff and then mold these into an operational system.

The model first set out to attempt to meet the following criteria:

1. It would be objective and not rely on judgments.
2. It would prevent data from being influenced or distorted by beliefs or biases of individual raters.
3. It would not be used to interpret or draw inferences.
4. It would be designed to eliminate personal involvement.

5. All behaviors would be clearly defined.
6. Behavior description scales would explicitly state what was to be observed and be observable.

In order to qualify as an observation model for evaluation of instructional personnel that would be practical and useful to school districts, information from the literature review was used to establish that the model should be:

1. Descriptive
2. Objective
3. In a format that could be easily learned and mastered with a minimum of effort and training
4. Manageable by administrative personnel
5. Useful to the classroom teacher.

The model establishes classroom observation procedures. A team of three observers were to observe and record teaching behaviors based upon scales to be developed in each of the three field sites. Following the classroom observation, the observers recorded data would be analyzed. The observers would then conduct a conference with the teacher to share their recorded observations. At this time an action plan for improvement could be developed or a second observation could be scheduled. Observers would stop at any step in the process when insufficient results were indicated. They would then return to a previous step to obtain the necessary information to proceed and complete the process. Satisfactory completion of the process provides for looping back to

the first step, classroom observation, and following the process again. See Figure 1.

Population for Field Testing

Districts where previous working relationships had been established were first contacted to solicit their cooperation in serving as field sites. Several school districts close to Des Moines were unable to participate because of constraints in their master contracts with teachers. The three school districts that did agree to participate were selected because of their interest and commitment to the project. They also did not have the constraints of a master contract. The three school districts selected were located in two states and represented three different demographic regions. Two of the school districts were urban and one was a rural school district. Each district's key contact person was a personal acquaintance from previous contacts.

One field test site used the model for special education personnel and the other two sites used general education instructional personnel for collecting performance evaluation data. No assumption was made that there is a difference between school size or urban compared to rural. Three field sites did provide more data regarding observer interactions and training than one site would have.

There were four secondary principals, twelve elementary principals, three special education administrators, and one special education supervisor participating in the

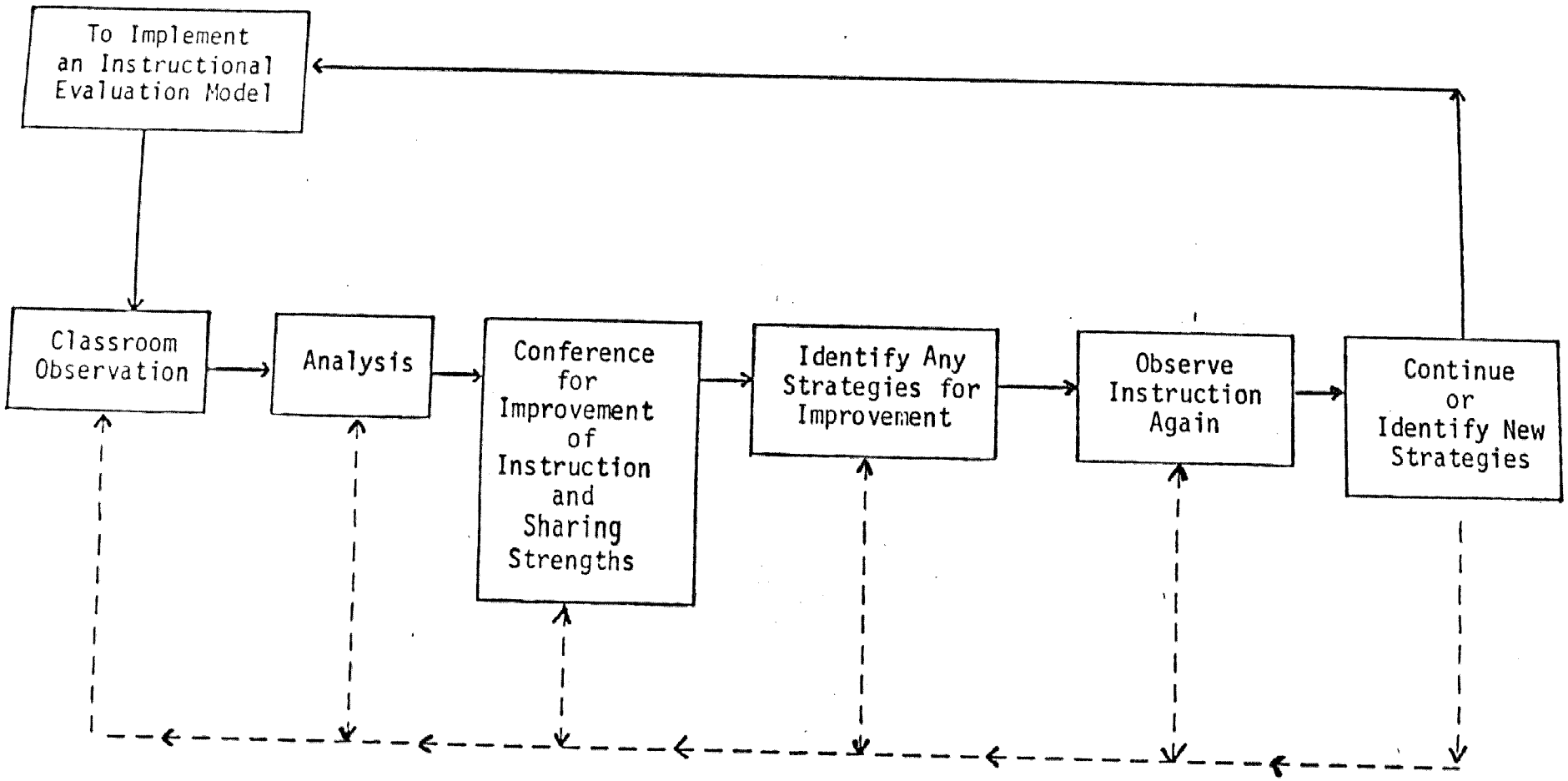

Figure 1

A Process for the Evaluation Model

field testing. The student enrollment of the three districts ranged from a low of 1,390 students to a high of 11,600 students. The school settings where observations were conducted for field testing the model ranged from a rural unincorporated village to a metropolitan suburban community.

The principals and other school administrators from the field test sites selected the teachers to be observed for field testing the model. The teachers selected had a range in competencies as perceived by the administrators.

A total of twenty-two teachers were selected and observed by the teams for collecting data on the model. Table 1 represents the demographic data of the teachers observed by the field site teams.

Developing the Performance Evaluation Scales

In the development of the model for evaluating instructional performance, the applicability of the local situation was a factor that had to be considered. Therefore, the instrument for evaluation had to reflect the local district's goals and philosophy.

It was determined the district goals could assist in the development of the evaluation plan and to provide direction. Once the goals of the district were established, the decision as to which procedures, behavioral scales and other techniques were to be incorporated into the evaluation system were addressed. Once the goals were identified or

Table 1

Demographic Data of Teachers Observed in the Three Field Sites for Training and Data Collection

Grades Taught	No.	Age	Sex M	Sex F	Years Experience
K-3	13	29-46		13	4 to 14
4-6	2	32-42		2	7 to 14
7-9					
10-12	2	25-41	2		4 to 20
Jr. Hi. EMH*	1	41	1		13
Jr. Hi. MC**	1	25		1	3
Sr. Hi. BI***	2	26-35		2	3 to 11
Sr. Hi. EMH****	1	55	1		20

* Junior high educable mentally handicapped
** Junior high multi-categorical
*** Senior high behaviorally impaired
**** Senior high educable mentally handicapped

developed, these served as the standards for selecting the instructional behaviors.

The approach was the development of a model for school districts to use for evaluating instructional performance that is observable in the classroom. Because it was task and function oriented, the model provided a natural framework from which to observe teaching behaviors.

Initially, principals were asked to list and describe teaching behaviors that they thought teachers needed to be successful in the classroom. Next, these events were shared with classroom teachers for their additions and modifications. These events were then collapsed into a workable number which represented those with commonality and most relevant to the district goals. This procedure has been supported by Grant and Carvel.[1] They conducted a study to determine whether teachers and principals agreed on what constitutes desirable and undesirable teaching behaviors. The data they gathered supported that there was a high degree of agreement between principals and teachers concerning teacher evaluation criteria.

In the initial development and later revisions of the behavior description scales, a list was circulated among the administrative and/or instructional staff in the field test

---

[1]Stephen Grant and Robert Carvel, "A Survey of Elementary School Principals and Teachers: Teacher Evaluation Criteria," Education, 100 (Spring 1980), 223-26.

sites to look for:

1. Comprehensiveness. To check for omissions in areas of teaching and to add any missing.
2. Selectivity. Statements that might be irrelevant or unimportant were deleted.
3. Priority. To consider the most important items of instructional behaviors.

On the basis of the faculty generated statements, eight to twelve behavior categories were selected by the field test site districts. See Appendix C for sample observation scales. The behavior statements were generally retained for the behavior scale titles. In cases where more than one statement was combined, the more descriptive statement was used or a new statement was written. Under each behavior scale title, a definition was given for that particular title. The definition established the baseline for what the given statement meant for that school district's purpose of assessing instructional performance. A behavior description scale was then written for each title statement. The behavior descriptions provided the clues and content for what raters were to observe in the classroom. Three statements, referred to as a scalogram, describing teaching behavior from the most desirable to the least desirable were listed to complete each behavior scale. In the final form the scalogram under each behavior description scale was scrambled to require the observer to relate the recorded

data to the behavioral terms and not make a choice based upon the position or weight of the scalogram.

The behavior statements generated by personnel of the field test sites were knowledge competencies of teaching process and not subject matter content. Since these knowledge competencies of the process of teaching were measured indirectly through observation methods, knowledge of the behavior is inferred from one's ability to perform it. To guard against the inference that a teacher could not perform a behavior because it was not observed, the scalogram also contained a statement of "not observed." Therefore, the observer could inquire about a particular behavior in the post observation conference, or make a notation to observe again for that particular behavior, but the observer was not to score any behavior not observed.

This model and the subsequent behavior description scales developed by the field test sites formed the instrument for behavioral observation performance of instructional personnel as locally defined. The observation instruments varied in length and written behaviors from one field site to another. This was primarily because of the different emphasis and priorities one district had that another one did not. It also was a direct reflection of the behavior description scales being related to a given district's goals.

The model for behavioral observation performance of instructional personnel states in behavioral terms what is

expected of the classroom teacher. The intent was that through classroom observation and interviews, it would be possible to ascertain with considerable accuracy a profile of a teaching performance on eight to twelve behavior scales.

There were several reasons for selecting teaching behaviors as the standard for the evaluation model. These included: reliable instruments for measuring teacher behaviors can be developed, use of teacher behaviors is based on the research, the strength of using teacher behaviors was demonstrated in the literature review and behavior can be learned by teachers.

Training of Observers

To increase the likelihood that an evaluation system would be developed with quality, the three field test site observers were required to receive training. Training of observers became a major component of the project. The training consisted of a thorough orientation of the district's goals and of the behavior description scales developed by the field sites. The observers were also provided a two and a half to three hour inservice on observation techniques, recording data and interviewing techniques. The training contained an activity based session for the observers to practice identifying observable, process, and characteristic behaviors.

Observers were first given an orientation to the major purpose of a performance appraisal system. Secondly, the

observers received training on the criteria for observing, including actions that should occur before, during and following an observation. A simulation activity was then conducted for the observers to become aware of the appropriate and inappropriate techniques for an actual classroom observation. This was followed by an activity based session having the observers practice identifying behaviors that could be observed, i.e., those that could be seen, heard, counted or measured. An answer sheet was provided with the behaviors that could be observed to stimulate discussion. The observers then received training on appropriate preparations for conducting the interview procedure.

The training materials corresponding to these steps are incorporated in Appendix A which includes: Objectives for Performance Appraisal, Observation Approach, Observation Techniques, Behavior Checklist and Performance Review Process.

The training session was followed with a video tape of classroom instructional activities. This not only served as part of the training exercise, but also allowed the principals and supervisors to become familiar with the observation and scoring forms prior to data collection.

Field Testing the Model

In the classroom data collection sessions, each observer independently observed and recorded their own data. They also independently scored their observations and only

compared their ratings after all observations were finished.

A post observation conference was conducted, the same day or the next day at the latest, to review the results with the instructional personnel being observed. The immediacy of the post observation conference was to ensure the rater would recall the meaning of his or her notes.

Significance of Observation State

For field testing the model, the teacher's actual classroom behavior was observed and recorded by only those who participated in the training and practice sessions. A teacher's actual and usual classroom procedures and techniques of teaching were observed and recorded on the field site's instrument developed according to the model. The behavior description categories focused the observers attention to low inference behaviors. Teaching behaviors, as identified on the behavioral descriptions scales, were observed and recorded as they appeared. Thus, a record was being provided to determine frequency or factual count feedback for the teacher.

The task of the observer was to observe activities that occurred in the classroom and then to record them on the observation recording form. See Appendix B. The observer was to make no attempt to score any observed behaviors during the observation session. The observer's crucial function was to serve as an abstractor, to select those behaviors

relevant to a particular behavior scale description and record them. The observers were to code the behaviors as they were observed, using their own style of notes and shorthand. Various techniques the observers could use for recording behaviors were tallying, checks, notes, counts, and other marks that would yield information about the behaviors which occurred. Only actual observed behaviors were to be recorded during the scheduled observation period. The observers were instructed on the importance of recording the behavior as soon as it was observed.

The behavioral categories that the observers used had been defined for both the teacher and the observer.

The prespecified observation behavior description scales permitted comparison between the field site observers, which provided the data for determining the inter-observer or inter-rater reliability of the model and instrument.

Summation of Method

The systematic observation model as developed and applied in this project had as the principal component: trained observers which recorded and rated specific classroom behaviors.

The primary purpose of field testing the model was to: (1) determine inter-observer agreement in using the model, (2) test the reliability of the instrument, and (3) provide a bases for school districts to use only behavioral description scales that can be reliably observed.

Table 2 provides a visual display of the steps for the systematic observation model.

Table 2

Steps in Selecting Teacher Performance Behaviors to be Observed

Step I	Develop or secure school district's educational goals Review existing list Select top priority goals Administrative and instructional personnel generate behavior statements related to the goals Behavior statements reduced to priorities and workable number (8-12) (Systematic observation procedure established)
Step II	Begin building the observation instrument Define each behavior statement Write behavior description scales for each behavior statement Write scalogram for each behavior statement (Systematic evaluation procedures specified)
Step III	Develop procedures for training observers Training on observation techniques Practice on observable behaviors Practice use of instrument as a group using a videotape Practice use of instrument, teams of two to three in classroom (Model field tested)
Step IV	Collection of data Processing data (Data collection complete)
Step V	Results summarized Results to participating districts (Action results)

CHAPTER FOUR

Analysis of Data

This chapter presents an analysis of the data representing the degree of relationship between independent raters when observing classroom teaching behaviors. The data provides the level of support for the statement of the problem. The project addressed the problem: can the model reduce subjectivity and bias with high inter-rater reliability and have at least an 85 percent agreement between raters.

The project was designed to develop a performance evaluation model of instructional personnel and then tested for inter-rater reliability and percentage of agreement of raters. Specifically designed versions of the model for each field site were used for this purpose.

A Pearson product-moment correlation of coefficient was used to determine the project's major problem, that of reliability. The Statistical Package for Social Sciences and the TUSTAT Pearson correlation procedure were applied to determine the reliability coefficients. The level of significance for the training sessions was set at the .05 level for determining whether a significant relationship existed between raters. A .90 correlation coefficient was

established for the actual data collection sessions as significant. Many statisticians establish .80 as a substantial positive correlation. Bloomers and Linquist, however, recommend that to be absolutely sure, a .90 correlation should be used for significance prediction.[1]

Bloomers and Linquist also discuss that there is not a pure description or definition that is acceptable regarding what is meant by closeness or degree of relationships.[2] Many other measures could be used, but according to Bloomers and Lindquist, they are generally not as convenient as the correlation of coefficient and it is more preferable.[3] For determining the level of significance for percentage of agreement between raters, an 85 percent of agreement was established. This figure is common and supported in the literature as an appropriate level for which inter-rater agreement should occur on performance evaluation instruments. Tables 3 through 36 present data results from the field site observations.

Two tables were developed for each of the three training sessions and the nineteen classroom observations. One table represents the correlation coefficient analysis and the

---

[1]Paul Bloomers and E. F. Lindquist, Statistical Methods in Psychology and Education (Boston: Houghton Mifflin, 1960), pp. 400-406.

[2]Ibid.

[3]Ibid.

second table displays the percentage of agreement between raters for each training and data collection session. The percentage of agreement and reliability tables are presented in sequential order for field sites A through C. The mean percentage of agreement for all raters was also computed for each observation and training session.

Table 3

Reliability of Eight Observers Viewing a 30 Minute Video Tape of a First Grade Class Training Session - Field Site A

Observer	1	2	3	4	5	6	7	8
1	X	.28	-.25	.72*	.55	.63*	.21	.59*
2		X	.57*	-.03	.48	.38	.86*	-.32
3			X	-.36	.32	.25	.66*	-.53
4				X	.38	.05	-.17	.71*
5					X	.32	.32	.44
6						X	.48	-.05
7							X	-.48
8								X

SPSS Pearson correlation matrix
* Significant at the .05 level

Reliability coefficients at the .05 level of significance was reached on only seven of twenty-eight possible interactions. Correlations ranged from negative to positive. Two observers in the training session were not a part of

the field site's data collection team. They participated in the training exercise for interest and support of the data collection that was to occur in the field site. One of the participants was the superintendent of schools.

Table 4

Inter-observer Agreement of Eight Observers Viewing a 30 Minute Video Tape of a First Grade Class Training Session - Field Site A

Observer	1	2	3	4	5	6	7	8
1	X	71	79	68	94*	94*	88*	85*
2		X	90*	95*	81	81	81	79
3			X	86*	84	84	89*	80
4				X	73	73	77	89*
5					X	100*	94*	84
6						X	94*	76
7							X	77
8								X

Mean group percent of agreement = 83%
*85% agreement level

Observers in site A did not reach an 85 percent agreement on the training session. The percentage of agreement, however, was only two percentage points less than that established as acceptable. Percentage of agreement ranged from 68 percent to 100 percent, with twelve agreements of 85 percent or higher. Two of the observers were interested

central office personnel and did not intend to participate in the subsequent field testing sessions.

Table 5

Reliability of Three Observers During a Second Grade Phonics Class Data Collection Session - Field Site A

Observer	1	2	3
1	X	.97	.96
2		X	.98
3			X

TUSTAT Pearson correlation matrix

Table 6

Inter-observer Agreement of Three Observers During a Second Grade Phonics Class Data Collection Session - Field Site A

Observer	1	2	3
1	X	96	90
2		X	87
3			X

Mean group percent of agreement = 91%

The reliability coefficients of all three observers were above a .90 correlation. Percentage of agreements ranged from 87 percent to 96 percent with a mean of 91 percent.

percentage of agreement between observers all surpassed the 85 percent level of acceptance. The inter-rater reliability coefficients were all above the levels determined to be statistically significant.

Table 7

Reliability of Two Observers During an Eleventh Grade German Class Data Collection Session - Field Site A

Observer	1	2
1	X	.97
2		X

TUSTAT Pearson correlation matrix

Table 8

Inter-observer Agreement of Two Observers During an Eleventh Grade German Class Data Collection Session - Field Site A

Observer	1	2
1	X	92
2		X

Mean group percent of agreement = 92%

The reliability coefficient of the two observers was .97 and a 92 percent of agreement was obtained. Both the correlation and percent of agreement exceed the

pre-established levels. A near perfect correlation was obtained on the observation and the percent of agreement was several percentage points above 85 percent. Nearly every observation scale was rated identical by each observer.

Table 9

Reliability of Four Observers During a First Grade Reading Class Data Collection Session - Field Site A

Observer	1	2	3	4
1	X	.95	.96	.93
2		X	.97	.94
3			X	.99
4				X

TUSTAT Pearson correlation matrix

Table 10

Inter-observer Agreement of Four Observers During a First Grade Reading Class Data Collection Session - Field Site A

Observer	1	2	3	4
1	X	97	97	93
2		X	100	96
3			X	96
4				X

Mean group percent of agreement = 96.5%

Reliability coefficients of .93 to .99 was obtained by the four observers of a first grade reading class. All correlations were above the .90 level for significance. Their percent of agreement ranged from 93 percent to 100 percent with a mean percent of 96.5.

Table 11

Reliability of Four Observers During a First Grade Math Class Data Collection Session - Field Site A

Observer	1	2	3	4
1	X	.94	.95	.97
2		X	.92	.94
3			X	.90
4				X

TUSTAT Pearson correlation matrix

Table 12

Inter-observer Agreement of Four Observers During a First Grade Math Class Data Collection Session - Field Site A

Observer	1	2	3	4
1	X	96	93	89
2		X	96	93
3			X	97
4				X

Mean group percent of agreement = 94%

The four observers of the first grade math session had reliability coefficients from .90 to .97. A mean percent of agreement between the observers was 94 percent with agreement between individual observers ranging from 89 to 97 percent. Both results exceed the established levels for acceptance.

Table 13

Reliability of Four Observers During a Fourth Grade Language Arts Class Data Collection Session - Field Site A

Observer	1	2	3	4
1	X	.94	.96	.95
2		X	.96	.93
3			X	.98
4				X

TUSTAT Pearson correlation matrix

Table 14

Inter-observer Agreement of Four Observers During a Fourth Grade Language Arts Class Data Collection Session - Field Site A

Observer	1	2	3	4
1	X	100	67	87
2		X	67	87
3			X	79
4				X

Mean group percent of agreement - 81%

Observers during a fourth grade language arts session obtained reliability coefficients of .93 to .98. The mean percent of agreement among these same four observers was, however, 81 percent. Although the correlations were all above the .90 significant level, the percent of agreement was four percentage points below the established level.

Table 15

Reliability of Four Observers Viewing a Video Tape of a Second and Third Grade Combined Class Training Session - Field Site B

Observer	1	2	3	4
1	X	.84	1.00	1.00
2		X	.84	.84
3			X	1.00
4				X

TUSTAT Pearson correlation matrix

Table 16

Inter-observer Agreement of Four Observers Viewing a 30 Minute Video Tape of a Second and Third Grade Combined Class Session Training Session - Field Site B

Observer	1	2	3	4
1	X	60	60	60
2		X	100	100
3			X	100
4				X

Mean group percent of agreement = 80%

Reliability coefficients ranged from .84 to 1.00. One-half of the correlations were below and one-half were above the .90 level. The training group had agreement from 60

percent to 100 percent, with a mean of 80 percent.

Table 17

Reliability of Four Observers During a First Grade Class Data Collection Session - Field Site B

Observer	1	2	3	4
1	X	1.00	1.00	.96
2		X	1.00	.96
3			X	.96
4				X

TUSTAT Pearson correlation matrix

Table 18

Inter-observer Agreement of Four Observers During a First Grade Class Data Collection Session - Field Site B

Observer	1	2	3	4
1	X	100	100	82
2		X	100	82
3			X	82
4				X

Mean group percent of agreement = 91%

Observer reliability coefficients of the four observers yielded correlations of .96 to 1.00, all above the .90

level for significance. Inter-observer agreement ranged from 82 percent to 100 percent with a mean agreement of 91 percent.

Table 19

Reliability of Three Observers During a Kindergarten Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 20

Inter-observer Agreement of Three Observers During a Kindergarten Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

Three observers during the kindergarten class data collection session all had a reliability coefficient of

1.00. They also all had a percent of agreement of 100 percent. These results yield the maximum that could be obtained from the observation data.

Table 21

Reliability of Three Observers During a Second Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 22

Inter-observer Agreement of Three Observers During a Second Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

The three observers' reliability coefficients during the second grade class data collection session were all a

1.00 correlation. Their inter-observer agreement level was also 100 percent. Perfect correlation and perfect agreement was reached among observers.

Table 23

Reliability of Three Observers During a Third Grade Class Data Collection Session – Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 24

Inter-observer Agreement of Three Observers During a Third Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

Reliability coefficients for the three observers during the third grade class data collection session had

1.00 correlations. A 100 percent agreement was also obtained between observers. The mean group percent of agreement was also 100 percent.

Table 25

Reliability of Four Observers During a Second Grade Class Data Collection Session - Field Site B

Observer	1	2	3	4
1	X	1.00	1.00	.95
2		X	1.00	.95
3			X	.95
4				X

TUSTAT Pearson correlation matrix

Table 26

Inter-observer Agreement of Four Observers During a Second Grade Class Data Collection Session - Field Site B

Observer	1	2	3	4
1	X	100	100	73
2		X	100	73
3			X	73
4				X

Mean group percent of agreement = 87%

Reliability coefficients ranged from .95 to 1.00. The mean percent of agreement among the observers was 87 percent. Both results were above the established levels for acceptance.

Table 27

Reliability of Three Observers During a First Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 28

Inter-observer Agreement of Three Observers During a First Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

All reliability coefficients were 1.00 for the three observers during the first grade class data collection session. A 100 percent level of agreement was also reached among all observers. Mean group percent of agreement was also 100 percent.

Table 29

Reliability of Three Observers During a First Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 30

Inter-observer Agreement of Three Observers During a First Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

Reliability coefficients of 1.00 were also reached for this first grade class data collection session by these three observers. Their percent of agreement was also 100 percent. Perfect correlations and agreement were reached between all observers.

Table 31

Reliability of Three Observers During a Fifth Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 32

Inter-observer Agreement of Three Observers During a Fifth Grade Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

For the fifth grade class data collection session the observers obtained reliability coefficients all at 1.00. Mean percent of agreement and agreement among observers were 100 percent.

Table 33

Reliability of Three Observers During a Second and Third Grade Combined Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	1.00	1.00
2		X	1.00
3			X

TUSTAT Pearson correlation matrix

Table 34

Inter-observer Agreement of Three Observers During a Second and Third Grade Combined Class Data Collection Session - Field Site B

Observer	1	2	3
1	X	100	100
2		X	100
3			X

Mean group percent of agreement = 100%

Reliability coefficients were all 1.00 for this second and third grade combined class data collection session. The percent of agreement was also 100 percent between all observers.

For these data and all the other tables yielding correlations of 1.00 and 100 percent of agreement, the raw data was retabulated and entered on the statistical package to ensure their accuracy. Results from the second application were identical to the first application. A contact was also made to the field site to verify that the raw data were accurate. All observers in the field site confirmed that these raw data were correct as submitted.

Table 35 provides analysis of the data from the training session conducted in field site C. Twenty-three correlations at the .05 level out of a possible ninety-one interactions were found. Table 36 shows that the mean group percent of agreement among observers for the training session in field site C was 80 percent, with percentages ranging from 41 percent to 100 percent. Six of the participants were not scheduled to perform data collection sessions. These six participants also moved in and out of the training session when it was being conducted. A brief review was provided for these individuals prior to observing the video tape. Trainees indicated that they found the physics lecture difficult to observe for a variety of teaching behaviors. Most of the trainees stated they only were able to observe two

Table 35

Reliability of Fourteen Observers Viewing a 30 Minute Video Tape of a Senior High Physics Lecture Training Session - Field Site C

Observer	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	X	.64*	.01	.55	.80*	.79*	.66*	-.02	.23	.75*	-.01	-.35	.00	.11
2		X	-.36	-.02	.19	.77*	-.50	-.50	.29	.65*	.09	-.10	-.52	-.53
3			X	.23	.44	.08	.10	.68*	.36	.32	.40	.57	.53	.66*
4				X	.82*	.35	-.25	.59	.30	-.09	.18	-.25	.64*	.73*
5					X	.52	-.46	.47	-.47	-.39	.28	-.08	.42	.66*
6						X	.68*	-.15	.23	.63*	.03	-.15	-.10	-.09
7							X	.46	-.11	.89*	.14	.35	.32	.02
8								X	.34	.63*	.46	.37	.92*	.86*
9									X	.00	.94*	.69*	.22	.41
10										X	.28	.48	.54	.28
11											X	.84*	.36	.42
12												X	.21	.22
13													X	.84*
14														X

SPSS Pearson correlation matrix
*Significant at the .05 level.

Table 36

Inter-observer Agreement of Fourteen Observers Viewing a 30 Minute Video Tape of a Senior High Physics Lecture Training Session - Field Site C

Observer	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	X	65	100*	65	88*	82	41	94*	47	59	53	59	94*	100*
2		X	65	100*	73	79	64	61	73	91*	82	91*	69	65
3			X	65	88*	82	41	94	47	59	53	59	94*	100*
4				X	73	79	64	61	73	91*	82	91*	69	65
5					X	93*	47	83	53	67	60	67	94	88*
6						X	50	78	57	71	64	71	88*	82
7							X	39	88*	70	78	70	44	41
8								X	44	56	50	56	89*	94*
9									X	80	89*	80	50	47
10										X	90*	100*	63	59
11											X	90*	56	53
12												X	63	59
13													X	94
14														X

Mean group percent of agreement = 80%
*85% agreement level

or three behaviors that related to the scales they were using for observation. Some of the trainees admitted they recorded some behaviors that they knew were not observable, but inferred they would occur if in the actual class setting.

Table 37

Reliability of Three Observers During a High School Educable Mentally Handicapped Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	.88	.93
2		X	.98
3			X

TUSTAT Pearson correlation matrix

Table 38

Inter-observer Agreement of Three Observers During a High School Educable Mentally Handicapped Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	69	73
2		X	94
3			X

Mean percent of agreement = 79%

Two reliability coefficients were above and one only slightly below the significant level. Percent of agreements had a substantial range with a mean agreement of 79 percent, falling below the level of acceptance.

Table 39

Reliability of Three Observers During a Behaviorally Impaired High School Itinerant Data Collection Session - Field Site C

Observer	1	2	3
1	X	.85	.77
2		X	.95
3			X

TUSTAT Pearson correlation matrix

Table 40

Inter-observer Agreement of Three Observers During a Behaviorally Impaired High School Itinerant Data Collection Session - Field Site C

Observer	1	2	3
1	X	75	78
2		X	96
3			X

Mean percent of agreement = 83%

Reliability coefficients ranged from .77 to .95. One correlation coefficient was near the .90 level, one above, and the third a fair distance from it. Percent of agreements also had a range, running from 75 percent to 96 percent. The mean percent of agreement was within two percentage points of the acceptable level.

Table 41

Reliability of Three Observers During a Junior High Behaviorally Impaired Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	.83	.76
2		X	.64
3			X

TUSTAT Pearson correlation matrix

Table 42

Inter-observer Agreement of Three Observers During a Junior High Behaviorally Impaired Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	91	91
2		X	100
3			X

Mean percent of agreement = 94%

None of the reliability coefficients reached the .90 level. One was close but the other two were not. A near perfect agreement of 94 percent was found among observers.

Table 43

Reliability of Three Observers During a Junior High Educable Mentally Handicapped Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	.94	.81
2		X	.91
3			X

TUSTAT Pearson correlation matrix

Table 44

Inter-observer Agreement of Three Observers During a Junior High Educable Mentally Handicapped Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	100	86
2		X	86
3			X

Mean percent of agreement = 90%

Two reliability coefficients were found to be above and one below the .90 significant level. All inter-observer

agreements were above 85 percent, with a mean agreement of 90 percent.

Table 45

Reliability of Three Observers During a Junior High Multi-categorical Handicapped Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	.87	.88
2		X	.98
3			X

TUSTAT Pearson correlation matrix

Table 46

Inter-observer Agreement of Three Observers During a Junior High Multi-categorical Handicapped Class Data Collection Session - Field Site C

Observer	1	2	3
1	X	100	93
2		X	93
3			X

Mean percent of agreement = 95%

All three reliability coefficients clustered around the .90 level. Two were very slightly below and one very

near a perfect correlation. The percent of agreements were also very high, one at the 100 percent level. All exceeded the 85 percent agreement.

This chapter has presented the results and statistical treatment of the data of the three training sessions and the nineteen data collection sessions. Percentages of agreement among raters ranged from 79 percent to 100 percent, with sixteen observations being above 85 percent. The nineteen data collection sessions yielded a range in correlation from .64 to 1.00 with sixteen observations falling above the .90 established level for significance. Data have been analyzed and limited only to observed sessions to allow a reasonably similar interpretation by any person reading the results.

The raw data on the observation recording forms were also quite informative as they provided patterns and sequences of instructional behaviors for the field sites participating in the project.

The main purpose of the field sites was to test the model under actual conditions. This was to determine if it would perform as desired with intended users. Table 47 displays the procedure for collecting the data.

Testing of the model was to ensure that the model would support the larger intent: showing the model would work in new settings with support for replication. The analysis of data was to determine support of the model's adaptability

to local school settings.

Table 47

Procedure for Observer Agreement and Reliability

	Inter-observer
Purpose	Determine consistency with other observers
When	Prior to formal evaluations--data collection in actual classroom settings
Medium	Video tape for practice sessions actual classroom settings for field testing the model
Unit of Analysis	Pearson product-moment correlation Inter-rater percentage of agreement

The conclusions, recommendations and implications of these findings are presented in chapter five.

CHAPTER FIVE

Summary, Conclusions, and Recommendations

The purpose of this chapter is to summarize the study, draw conclusions based upon the findings, and formulate recommendations.

Summary

A coefficient of correlation and a percentage of agreement between raters was calculated for each classroom observation session.

The analysis of data was performed by using the Statistical Package for Social Sciences and TUSTAT Pearson product-moment coefficient of correlation procedure, on Drake University's VAX computer system.

The analysis of inter-rater reliability of the model yielded relationships of significance with a correlation of .90 for sixteen of the nineteen data collection observations. Mean inter-rater agreement of 85 percent was also found on sixteen of the nineteen observations. Seven of the observations had a 100 percent agreement and a 1.00 correlation among observers. Three observations fell below the .90 correlation, one with correlations of .64 to .83 and the other two had correlations of .77 and higher. The three

observations that were less than 85 percent agreement, were only slightly lower, one was 79 percent and two were 83 percent.

When the three field site results are viewed in relationship to each other, they are quite similar. The similarities of results would seem markedly close, especially when one considers the different populations used for data collection. Results of the data suggest that the performance evaluation model measured the degree to which the observers identified observable behaviors. The performance evaluation model yielded a quantitative record of what occurred in specific classrooms under specific conditions. The analysis suggests that objectivity and reliability, both necessary conditions for validity, were obtained.

Unsolicited comments by teachers were received by the observers and are shared in the findings. A sampling of the comments follows:

> Beneficial because an understanding was developed between administrator and teacher of what was expected.
>
> It helped my teaching and planning.
>
> It allowed me to focus on specific aspects of performance.
>
> Both teachers and administrators were aware of the same thing expected from the evaluation.
>
> Feedback was beneficial, objective and useful for me to plan.
>
> The first time evaluation has been something more than just an exercise.
>
> I wished that all teacher evaluations were like this.

A summary of the major findings of the project were that the model:

1. provided a common set of procedures and terms with the same meaning for teacher and administrator.

2. provided for increased awareness of classroom behaviors on behalf of the instructional personnel.

3. brought attention to desired instructional behaviors for meeting district goals.

Conclusions

The following conclusions have been drawn from the project:

1. Instructional personnel can be objectively and reliably observed in a classroom setting.

2. Principals can be trained to observe and accurately identify teacher behaviors.

3. Team leaders and supervisors for special education can be trained to observe and accurately identify teacher behaviors.

4. Previous studies of behavioral observations have been supported by this project.

The use of a specific set of pre-defined behavioral scale descriptions seem to benefit the observers by improving their observation skills. Pre-specified observation scale descriptions permitted comparison among the various observers for determining the consistency with which different raters observing the same teacher recorded the same or similar behaviors.

It should be recognized the behavioral description scales can only be a sampling of the knowledge process

behaviors. This same limitation would exist in any performance evaluation instrument, unless it were so comprehensive it would be rendered severely impractical. The real question is whether the behavior description scales are a sampling of the district's standards and the ones most in need of evaluation. Since the behavioral description scales were selected by the local school district according to philosophy and need, it is accepted that for their purpose it is an accurate sampling. The review of literature states that the use of school personnel in the selection of the behavior statements helps assure local district validity.

It seems evident that the model is sufficient for use by school districts interested in a performance evaluation system that relates directly to their local situation. It also seems evident that training of observers is essential to secure high reliability. The evidence appears to support that a high level of reliability exists through the use of the model.

The data supports the performance evaluation model as an accepted method of organizing observed instructional behaviors into a procedure which allows trained administrators to observe, record, and analyze behaviors.

The ultimate value of the performance evaluation model is the use which administrators and instructional personnel will make of the objective results from the observation sessions.

The performance evaluation model, if followed according to the procedures described in this project, will reflect what is going on in a classroom, as observed by a trained evaluator. Through the post-evaluation conference, teachers may be provided the opportunity to become more sensitive to their instructional behaviors and how these affect their classroom.

It is difficult to draw conclusions on how much performance evaluation models of the type in this project will be adapted in the future. There was no evidence to suggest that this model could not be readily adapted to specified goals of any district or unique priorities of a community and expect similar results as obtained in this project.

The efficacy of observational evaluation techniques is not yet fully known. By reporting the contents of this project, clarification of the similarities and differences of various approaches have been made. Perhaps some of the issues have been resolved that heretofore have prevented educators from using observational evaluation techniques.

Schools continue to change through the economic, social and political realities of the community where the school district is located. It would seem a necessity therefore, that any performance evaluation model be related to the educational goals of the community. School personnel and the school community will have to be alert to these changes, modifying their performance evaluation system accordingly.

No one has probably made this case better than Alvin Toffler:

> Education . . . is not just something that happens in the head. It involves our muscles, our senses, our hormonal defenses, our total biochemistry. Nor does it occur solely within the individual. Education springs from the interplay between the individual and a changing environment. The movement to heighten future consciousness in education, therefore, must be seen as one step toward a deep restructuring of the links between schools, colleges, universities and the communities that surround them.[1]

Recommendations

The project has added descriptive information on how to develop and implement a systematic observation system. This evaluation system proved objective and reliable within the parameters of present research. The information obtained from the project provides implications for further investigation. It is recommended that the behavioral description scales used in the field sites be replicated in other districts to determine their replicability and reliability.

It is further recommended that:

1. Follow-up with teachers should occur to elicit the degree of support they have for the model.
2. Reactions from the users be formally recorded as to their satisfaction.
3. Instructional personnel be included in any design or modifications of the performance evaluation model.

---

[1]Alvin Toffler, Learning for Tomorrow: The Role of the Future in Education (New York: Vintage Books, 1974), p. 13.

4. A five point, rather than a three point, scalogram be developed for each behavior scale for more accurate discrimination and matching of the observed teaching behaviors.

5. The systematic observation procedures be packaged as a self-instructional product for school administrators to use.

6. Both administrators and teachers be informed of the purpose at the beginning of the process for evaluation.

7. Acceptance and use of any behavior description scale only occur after inter-rater reliability has been established.

8. The stability of the model be tested by observing the same teacher with different pupils, settings and/or curriculum content.

9. The model be tested with observers outside the school district to ascertain the same high degree of reliability.

Comments

The project suggests a reliable practice of evaluation procedures has been demonstrated. Accurate and objective observations would seem to improve communication and respect between administrators and teachers. The complexity of any classroom defies many attempts for easy analysis and evaluation. This observation evaluation model seems to have the potential for simplifying this complexity.

The results of the project are encouraging for two reasons. Inter-observer agreements, after training in the use of the model and instruments, were all very high. The inter-observer agreements were all above the 85 percent level except for three observation sessions. Coefficient of

correlations were .90 on sixteen of the nineteen observations, which established high reliability. Another cause for encouragement is the evidence from the data and from observer comments, that the behavior scale descriptors reduced the ambiguity of recording and scoring teaching behaviors.

Documentation of the essential elements of the model have been provided to facilitate replication with as few complications as possible. The evidence of the effectiveness of the model has been reported with descriptions and samples of the materials used in the project.

In summary, similarities of inter-observer agreement and coefficient of correlations seem strong enough to suggest that differently developed instruments based upon the model's procedures does not affect the reliability of the model.

BIBLIOGRAPHY

BIBLIOGRAPHY

Books

Barr, Arvil S. Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness: A Summary of Investigations. Madison, Wisconsin: Dembar Publications, 1961.

Bedeian, Arthur. Organizations: Theory and Analysis. Hinsdale, Illinois: The Dryden Press, 1980.

Beecher, Russell S. Staff Evaluation: The Essential Administrative Task. Bloomington, Indiana: Phi Delta Kappa Educational Foundation, 1979.

Bellack, Arno A., et al. The Language of the Classroom. New York: Teachers College Press, Columbia University, 1966.

Bishop, Leslie J. "Systems for Observing In-School Operations." In Observation Methods in the Classroom. Eds. Charles Beagle and Richard Brandt. Washington, D.C.: Association for Supervision and Curriculum Development, 1973.

Bloomers, Paul, and E. F. Lindquist. Statistical Methods in Psychology and Education. Boston: Houghton Mifflin, 1960.

Borg, Walter R. Applying Educational Research. New York: Longman, 1981.

Borich, Gary D., ed. The Appraisal of Teaching: Concepts and Process. Reading, Massachusetts: Addison-Wesley, 1977.

_____, and Susan K. Madden. Evaluating Classroom Instruction: A Sourcebook of Instruments. Reading, Massachusetts: Addison-Wesley, 1977.

Brandt, Richard. "Toward a Taxonomy of Observational Information." In Observation Methods in the Classroom. Eds. Charles Beagle and Richard Brandt. Washington, D.C.: Association of Supervision and Curriculum Development, 1973.

Brophy, Jere E. Stability in Teacher Effectiveness. Austin, Texas: The Research and Development Center for Teacher Education, University of Texas, 1972.

Coleman, Peter. "The Improvement of Aggregate Teaching in Effectiveness in a School District." In The Appraisal of Teaching: Concepts and Process. Ed. Gary D. Borich. Reading, Massachusetts: Addison-Wesley, 1977.

Cummings, L. L., and Donald P. Schwab. Performance in Organizations. Glenview, Illinois: Scott, Foresman,

Dunn, Kenneth, and Rita Dunn. Administrators Guide to New Programs for Faculty Management and Evaluation. West Nyack, New York: Parker Publishing, 1977.

Gallagher, James J., Graham A. Nuthall, and Barak Rosenshine. Classroom Observation. Chicago: Rand McNally, 1970.

Glass, Gene V. "A Review of Three Methods of Determining Teacher Effectiveness." In The Appraisal of Teaching: Concepts and Process. Ed. Gary D. Borich. Reading, Massachusetts: Addison-Wesley, 1977.

Griffith, Francis. A Handbook for the Observation of Teaching and Learning. Midland, Michigan: Pendell Publishing, 1973.

Herman, Jerry J. Developing an Effective School Staff Evaluation Program. West Nyack, N.Y.: Parker Publishing, 1973.

Hyman, Ronald T. School Administrators Handbook of Teacher Evaluation Methods. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

_____. School Administrators Handbook of Teacher Supervision and Evaluation Methods. Englewood Cliffs, New Jersey: Prentice-Hall, 1981.

Klein, Stephen, and Marvin C. Alkin. "Evaluating Teachers for Outcome Accountability." In The Appraisal of Teaching: Concepts and Process. Ed. Gary D. Borich. Reading, Massachusetts: Addison-Wesley, 1977.

Medley, Donald M., and Harold E. Mitzel. "Measuring Classroom Behavior by Systematic Observation." In Handbook of Research on Teaching. Ed. N. L. Gage. Chicago: Rand McNally, 1963.

Millman, Jason. Handbook of Teacher Evaluation. Beverly Hills, California: Sage Publications, 1981.

Morrison, Arnold, and Donald McIntyre. Teachers and Teaching. Baltimore: Penguin Books, 1969.

Ober, Richard L., Ernest L. Bentley, and Edith Miller. Systematic Observation in Teaching. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Popham, W. James. Educational Evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Roettger, Walter B. Performance Appraisal Skills for Managers and Supervisors. Des Moines, Iowa: Institute of Public Affairs and Administration, Drake University, 1981.

Rosenshine, Barak, and Norma Furst. "The Use of Direct Observation to Study Teaching." In Second Handbook of Research on Teaching. Ed. Robert M. W. Travers. Chicago: Rand McNally, 1973.

Scriven, Michael. "The Evaluation of Teachers and Teaching." In The Appraisal of Teaching: Concepts and Process. Ed. Gary D. Borich. Reading, Massachusetts: Addison-Wesley, 1977.

Sirotnik, Kenneth A. "An Inter-Observer Reliability Study of the SRI Observation System as Modified for Use in a Study of Schooling." In A Study of Schooling. Los Angeles, California: University of California, Technical Report No. 27, 1981.

Soar, Robert S. "An Integration of Findings from Four Studies of Teacher Effectiveness." In The Appraisal of Teaching: Concepts and Process. Ed. Gary D. Borich. Reading, Massachusetts: Addison-Wesley, 1977.

Six Areas of Teacher Competence. Burlingame, California: California Teachers Association, 1964.

Thomas, M. Donald. Performance Evaluation of Educational Personnel. Bloomington, Indiana: Phi Delta Kappa Educational Foundation, 1979.

Toffler, Alvin. Learning for Tomorrow: The Role of the Future in Education. New York: Vintage Books, 1974.

Walberg, Herbert J. Evaluating Educational Performance: A Sourcebook of Methods, Instruments, and Examples. Berkeley, California: McCutchan Publishers, 1974.

Van Dalen, Deobold B. Understanding Educational Research. New York: McGraw-Hill, 1973.

Periodicals

Borich, Gary D. David Molitz and Cherry L. Kugle. "Convergent and Discriminant Validity of Five Classroom Observation Systems: Testing a Model." Journal of Educational Psychology, 70 (April 1978), 119-28.

Burke, Ronald J., and Douglas S. Wilcox. "Characteristics of Effective Employee Performance Review and Development Interviews." Personnel Psychology, 22 (Autumn 1969), 291-305.

Caldwell, Peggy. "Teacher-Evaluation Methods Called Inadequate." Education Week, 1 (November 1981), 4.

Ellman, Neil. "Evaluating Representative Teacher Behavior." National Association of Secondary School Principals Bulletin, 60 (September 1976), 25-27.

Enns, T. "Rating Teacher Effectiveness: The Functions of the Principal." The Journal of Educational Administration, 3 (March 1965), 81-95.

Goodlad, John. "Educational Leadership: Toward the Third Era." Educational Leadership, 35 (January 1978), 322-31.

Frick, Ted, and Melvyn I. Semmel. "Observer Agreement and Reliabilities of Classroom Observation Measures." Review of Educational Research, 48 (Winter 1978), 157-84.

Grant, Stephen, and Robert Carvell. "A Survey of Elementary School Principals and Teachers: Teacher Evaluation Criteria." Education, 100 (Spring 1980), 223-36.

Greer, Peter R. "Another Simple Truth." Education Week, 1 (June 1982), 20-24.

Heard, Alex. "N.C. to Begin Statewide Evaluation of Teachers, Principals." Education Week, 25 (August 1982), 6.

Herman, Jerry J. "Developing a Staff Evaluation Program." National Association of Secondary School Principals Bulletin, 60 (September 1976), 8-14.

Kaye, Beverly L., and Shelly Krantz. "Preparing Employees: The Missing Link on Performance Appraisal Training." Personnel, 59 (May-June 1982), 23-29.

Knezevich, Stephen J. "Designing Performance Appraisal Systems." New Directions for Education, 1 (Fall 1973), 37-50.

Kult, Lawrence C. "Improving Teacher Evaluation by Principals." The Clearinghouse, 52 (September 1978), 17-21.

Lamb, Morris L., and Kevin Swick. "A Historical Overview of Classroom Teacher Observation." Educational Forum, 39 (January 1975), 239-47.

Levin, Benjy. "Teacher Evaluation--A Review of Research." Educational Leadership, 37 (December 1979), 240-45.

Marks, Merle B. "Effective Teacher Evaluation." National Association of Secondary School Principals Bulletin, 60 (September 1976), 1-7.

McMillan, John D., and Hoyt W. Doyle. "Performance Appraisal: Match the Tool to the Task." Personnel, 57 (July-August 1980), 2-3.

Morrison, Ann, and Mary Ellen Krantz. "The Shape of Performance Appraisal in the Coming Decade." Personnel, 58 (July-August 1981), 12-22.

Robinson, John J. "The Observation Report--A Help or a Nuisance?" National Association of Secondary School Principals Bulletin, 62 (December 1978), 22-26.

_____, and John H. Lee, Jr. "Evaluation: Can We Agree?" National Association of Secondary School Principals Bulletin, 62 (December 1978), 15-20.

Rowe, Mary B. "Wait, Time and Rewards as Instructional Variables." Journal of Research in Science Teaching, 11, No. 2 (1974), 81-94.

Roy, Joseph J. "Teacher Evaluation in an Era of Educational Change." The Clearinghouse, 52 (February 1979), 275-76.

Schneider, Craig E., and Richard W. Beatty. "Performance Appraisal Revisited: Integrating Behaviorally Based and Effectiveness Based Methods." The Personnel Administrator, 24 (July 1979), 65-68.

Shymansky, James A. "Assessing Teacher Performance in the Classroom: Pattern Analysis Applied to Interaction Data." Studies in Educational Evaluation, 4 (Summer 1978), 99-106.

ERIC

Herbert, John. "A Research Base for Accreditation of Teacher Preparation Programs." In Accreditation and Research Problems. Eds. John L. Burdin and Margaret T. Reagan. ERIC ED 050 021.

Kugle, C. L. Data Collection Procedures for the Evaluation of Teaching Program, Phase III. ERIC ED 170 340.

Medley, Donald M., and Russell A. Hill. Measurement Properties of Observation Schedules and Record. ERIC ED 185 089.

Mitsakos, Charles L., and Kenneth R. Siefert. Teacher Evaluation Programs. ERIC ED 182 828.

Smith, B. Othanel. Certification of Educational Personnel. ERIC ED 055 975.

Ward, Beatrice A. Assessment of Teacher Performance: What is Involved? What is the Cost? ERIC ED 177 150.

Dissertations

Cole, Charles C. "A Comparison of Two Methods of Teacher Evaluation." Ed.D. dissertation, North Texas State University, 1978.

Neuenfeldt, John C. "An Investigation of an Alternative Method of Evaluating Classroom Teaching." Ed.D. dissertation, New Mexico State University, 1978.

Oakes, Ernest H. "The Educational Concerns and Priorities of Selected Parents in the Council Bluffs Community School District." Ed.D. dissertation, Drake University, 1981.

Newspaper

"Mehlville Teachers Protest Using Rating Scale to Eliminate Staff." South County Journal, Mehlville, Missouri, March 24, 1982, p. 9.

Nonprint Sources

Egglebrek, Dave. Evaluation of School Instructional Programs: How Do You Do It? Cassette. Washington, D.C.: Educational Resource Information Center, April, 1981.

Other

A Suggested Administrative Evaluation Form. Des Moines: Iowa Department of Public Instruction, 1974.

Legislative Bill 259. Lincoln: State of Nebraska, 1982.

"Teacher Evaluation." A Legal Memorandum. Reston, Virginia: National Association of Secondary School Principals, 1978.

APPENDICES

APPENDIX A

TRAINING MATERIALS

OBJECTIVES FOR PERFORMANCE APPRAISAL

- Recruitment
- Selection
- Placement
- Development and Training
- Appropriate Use of Personnel
- Maintenance
- (litigation)

OBSERVATION APPROACH

BEFORE:

1. Contact the teacher and arrange for observation time.
2. Have teacher complete objectives.
3. Inform teacher of purpose of the observation.

DURING:

1. Keep conversation at minimum.
2. Record observations promptly.
3. Avoid generalizations.

AFTER:

1. Leave room quietly.
2. Arrange for feedback to teachers as soon as possible.
3. Feedback conference - Be prepared for plan of action.

OBSERVATION APPROACH

Before

I. Setting - Variables

A. Time of day

B. Other events that day

C. Teachers activities before/after observation

D. Length of observation time

During

II. Systematic Procedure - Not Haphazard

A. Observe only behaviors that can be:

1. Seen
2. Heard
3. Counted
4. Measured

B. Process Behavior Not Observable

Example: Shyness

Memory

C. Characteristic Behavior Cannot be Observed!

Example: Honesty

Truthfulness

D. Techniques for Recording

1. Decide
2. Practice
3. Select
   a. Checklists
   b. Behavior tallying-charting
   c. Shorthand

After

III. Summarize and Interpret Notes - Tallies - etc.

A. Soon!

OBSERVATION TECHNIQUES

Appropriate	Inappropriate
Teachers informed in advance of observation purposes and criteria.	Teachers not informed.
Observer knows teacher's objectives for session.	Observer not aware of objectives or instructional plan.
Observer aware of school and/or IEP instructional goals.	Observer not aware.
Observer as inobtrusive as possible.	Observer's presence creates an artificial or unnatural climate.
Observation focuses on important dimensions of instruction, and only those.	Observation haphazard, and/or attends to irrelevant details.
Initially, focus is on analysis of whole scene to determine areas of need or problems.	Initial observations focused on isolated or nit-picking details.
Observer keeps an accurate record of observations.	Off-the-cuff comments.
After need area(s) identified, observation is focused on that area.	Observation remains general.
Observation provides specific feedback as a basis for change.	Feedback vague and general. Statements are value judgments without descriptions of behavior.
Observation identifies strengths as well as weaknesses.	Observations identify only weaknesses.
Observer provides feedback after class.	Observer intervenes in class or provides feedback in front of class.

BEHAVIOR CHECKLIST

Circle those words or phrases that denote a behavior (something that can be seen or heard, counted or measured).

fidgets
realizes
pronounces
understands
is shy
coughs
says
likes
matches
prefers
draws
hesitates
is inattentive
cries
fears
throws
is impulsive
is depressed
whines
trusts
selects
suppresses
runs
anticipates
desires
recognizes
kicks
perceives
comprehends
remembers
is honest
wants attention
appraises
is jealous
feels competent
differentiates
laughs
is aware
learns
is hostile
complains
daydreams
conceptualizes
drums
writes
is suggestible
feels inadequate
smiles
wiggles
is anxious
yawns
dreams
evaluates
thinks
questions
hits
interrupts

ANSWER SHEET
for
Behavior Checklist

Circle those words or phrases that denote a behavior (something that can be seen or heard, counted or measured).

fidgets
realizes
pronounces
understands
is shy
coughs
says
likes
matches
prefers
draws
hesitates
is inattentive
cries
fears
throws
is impulsive
is depressed
whines
trusts
selects
suppresses
runs
anticipates
desires
recognizes
kicks
perceives
comprehends
remembers
is honest
wants attention
appraises
is jealous
feels competent
differentiates
laughs
is aware
learns
is hostile
complains
daydreams
conceptualizes
drums
writes
is suggestible
feels inadequate
smiles
wiggles
is anxious
yawns
dreams
evaluates
thinks
questions
hits
interrupts

PERFORMANCE REVIEW PROCESS

1. Be prepared for the meeting

2. State the purpose of the review and put the staff member at ease.

3. Facilitate discussion of performance related issues by:

   a. Active listening

   b. Use of paraphrasing techniques

   c. Effective use of silence

   d. Being honest

4. Minimize personal criticisms

5. Use probing questions

6. Conclude with a plan of action

PERFORMANCE REVIEW PROCESS

1. Have "entrance" questions make the interview comfortable. Do you? Have you?

2. Proceed into "developmental" questions. What do you do? How do you?

3. Be specific - - probe. Ask for examples of when? Where? How often? Who participated?

4. Do not pressure or ask threatening questions. Examples: Not, what kind of records do you keep? Instead - - "Do you keep written records? If the answer is yes, then probe for more specifics.

5. Questions should be clear and concise.

6. Listen carefully to answers, pick up from there.

7. Avoid questions that cannot be verified or are irrelevant.

8. Stick to specifics. Avoid philosophy or feeling questions. You can gain this insight through specific or follow-up questions.

9. For the Interview Scale - ask questions that relate to the specifics of the Scale and the Descriptions.

PERFORMANCE REVIEW PROCESS

1. Open Question

Places no restrictions on the length of the respondent's answer. Gives the respondent more latitude in interpreting the subject to be discussed.

"What is your reaction to the inservice training program?"

2. Closed Question

More specific and usually requires a shorter, more direct answer.

"Do you like having to participate in the inservice training sessions?"

One important principle related to the use of open or closed questions is that these types of questions tend to influence the length of the interviewee's responses. Open questions encourage the respondent to talk more, while closed questions discourage participation. Since one of the problems in most interviews is getting the interviewee to become freely involved and to participate in the interview, open questions are more likely to be used in the early part of the interview or at the introductions of each new topic area, while closed questions are used as follow-ups for the responses to open questions.

3. Probing Question

Encourages the respondent to elaborate on what he has been saying. Why and How are common probing questions.

"I see. Can you tell me more?"

4. Loaded Question

Stacks the deck by implying the desired answer. A question of this type can be very detrimental to an interview.

"Isn't your inservice group behind the others?"

5. Obvious Answer Question of Leading Question

By its phrasing implies the expected response.

"You wouldn't mind taking a college course, would you?"

APPENDIX B

BEHAVIORAL OBSERVATION SCALE FORMAT AND SAMPLE FORMS

OBSERVATION SCALE

NAME OF SCALE:

DEFINITION:

SCALE DESCRIPTION:

SCALOGRAM:

A.

B.

C.

N. Not Observed

(Basic format for developing the specific content for the behavioral scales.)

OBSERVATION FORM

This form is to be completed by the observer during the observation visit to the teacher's classroom. It will serve as points for discussion during the feedback conference following the observation period.

1. Interpersonal Skills	5. Utilization of Organizational Techniques	9. Effective Lesson Checking for Understanding
2. Classroom Control	6. Effective Lesson Anticipatory Set	10. Effective Lesson Guided and Independent Practice
3. Plan Effectively	7. Effective Lesson Objective and Instructional Input	
4. Physical/Learning Environment	8. Effective Lesson Modeling	

Teacher: _____ School: _____ Date: _____ Subject: _____

OBSERVATION TALLY SHEET

	A	B	C	N	Value
1. INTERPERSONAL SKILLS					
2. CLASSROOM CONTROL					
3. PLAN EFFECTIVELY					
4. PHYSICAL/LEARNING ENVIRONMENT					
5. UTILIZATION OF ORGANIZATIONAL TECHNIQUES					
6. EFFECTIVE LESSON (Anticipatory Set)					
7. EFFECTIVE LESSON (Objective and Instructional Input)					
8. EFFECTIVE LESSON (Modeling)					
9. EFFECTIVE LESSON (Checking for Understanding)					
10. EFFECTIVE LESSON (Guided and Independent Practice)					

Total _____

KEY TO OBSERVATION SCALES

		5	3	1
1.	Interpersonal skills	B	A	C
2.	Classroom control	A	B	C
3.	Plan effectively	A	C	B
4.	Physical/learning environment	B	C	A
5.	Utilization of organizational techniques	A	C	B
6.	Effective lesson - anticipatory set	A	C	B
7.	Effective lesson - objective and instructional input	A	C	B
8.	Effective lesson - modeling	A	C	B
9.	Effective lesson - checking for understanding	B	C	A
10.	Effective lesson - guided and independent practice	A	C	B

APPENDIX C

SAMPLE FIELD SITE OBSERVATION INSTRUMENT

OBSERVATION SCALE

NAME OF SCALE: INTERPERSONAL SKILLS

DEFINITION: The teacher demonstrates a sincere regard and mutual respect for students in a cooperative and natural relationship.

SCALE DESCRIPTION: The teacher provides a classroom exhibiting a friendly, cooperative and mutually helpful atmosphere to foster student positive self-concept.

SCALE DESCRIPTORS:

A. Accepts students but is largely concerned with subject matter.

B. Exhibits a friendly, cooperative and mutually helpful atmosphere to foster a positive self-concept.

C. Shows little concern for needs of students and atmosphere developed.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: CLASSROOM CONTROL

DEFINITION: The teacher demonstrates classroom control by creating a healthy and secure feeling of freedom in the classroom.

SCALE DESCRIPTION: Classroom control is an atmosphere where one observes industrious and gainful self-regulation by the students. What the teacher does to assist and maintain student self-direction, expression, and involvement as an individual is an indication of the effectiveness of the teacher's ability for class-control. The class should be involved in self-discipline, self-directed, goal oriented activities. Freedom and security should be maintained on a balanced scale for the students within the classroom.

SCALE DESCRIPTORS:

A. Encourages student cooperation in maintaining an atmosphere of gainful self-regulation and is alert to loss of class control. He/she also identifies and takes responsibility for discipline problems.

B. Imposes standards of conduct to maintain control.

C. Uses only an authoritarian method when maintaining classroom control.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: PLAN EFFECTIVELY

DEFINITION: The teacher will have identified major target objectives for a lesson and have located students' educational position in relation to those objectives to create an effective lesson.

SCALE DESCRIPTION: Planning is one of the most influential factors in successful teaching. There should be a system to planning. Within each general content area, the teacher will have determined the particular objectives for teaching the lesson. The teacher will have located students' educational position in relation to the objectives to be taught.

SCALE DESCRIPTORS:

A. Proficiency in the ability to identify major objectives to be taught in a lesson and uses appropriate techniques, materials and individualization with the children.

B. Seldom demonstrates preparation of the lesson.

C. Some preparation of lesson and uses of different techniques and materials with the children are apparent.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: PHYSICAL/LEARNING ENVIRONMENT

DEFINITION: The teacher uses the classroom facilities and equipment to favorably influence the learning environment.

SCALE DESCRIPTION: The classroom physical environment needs to be conducive to learning, or can fail to reinforce the learning situation. The appearance of a room reveals whether basic considerations have been given to physical comfort, light and heat control, utilization of equipment, placement of resources for learning and their relevancy to tasks at hand. The teacher should regularly assume these responsibilities in developing an environment appropriate and functional for effective learning opportunities.

SCALE DESCRIPTORS:

A. Indication that no real attention has been focused on room arrangement both for physical and learning factors.

B. Planning and arrangement of a setting that relates to current objectives and provides for a favorable learning environment is apparent.

C. Modification of room arrangement is apparent but is not closely related to learning objectives of the classroom.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: UTILIZATION OF ORGANIZATIONAL TECHNIQUES

DEFINITION: The teacher demonstrates organizational skills in the classroom.

SCALE DESCRIPTION: Organizational skills are an important aspect in the classroom. The teacher should have a complete set of lesson plans appropriate to subject matter that are ready and useable. Classes should begin promptly and the students should be familiar with classroom procedures. A daily routine should be apparent through student behavior.

SCALE DESCRIPTORS:

A. Utilizes organizational skills to benefit both classroom instuctions and students.

B. Seldom demonstrates organizational skills in classroom and daily routine.

C. Maintains some organization of classroom instruction and daily routine and procedures.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: EFFECTIVE LESSON (Anticipatory Test)

DEFINITION: The teacher elicits attending behavior (deliberate focus) and a mental readiness for instruction.

SCALE DESCRIPTION: The teacher develops an anticipatory set by focusing the students attention, provide a very brief practice on previously achieved and related learning, or develop a readiness for instruction to follow. The activity should continue only long enough to get students ready, allowing the major portion of instructional time for accomplishment of current instruction.

SCALE DESCRIPTORS:

A. Elicits attending behavior and mental readiness for accomplishment of current instruction.

B. Elicits no attending behavior and mental readiness with the children for the current instruction.

C. Elicits some attending behavior and mental readiness with the children for the current instruction.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: EFFECTIVE LESSON (Objective and Instructional Input)

DEFINITION: The teacher states the objective of the lesson to the students and has pre-determined the necessary skills to accomplish the objective.

SCALE DESCRIPTION: The teacher communicates to the student what he will be able to do by the end of the instruction and why that accomplishment is important, useful and relevant. EXAMPLE: "Today we are going to learn ways of participation in a discussion so we each get turns and learn from other peoples' ideals." The teacher has also taken steps to determine what information (new or already processed) is needed by the student in order to accomplish the objective. Often students' are expected to achieve an objective without having been taught that which is necessary in order to do so.

SCALE DESCRIPTORS:

A. Communicates to the students the objectives of the instruction and provides the necessary prerequisite skills to master those objectives.

B. Introduces new instruction and states no objectives nor develops prerequisite skills.

C. Introduces the lesson and does not state instructional objectives; however, the necessary prerequisite skills to master objectives are provided.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: EFFECTIVE LESSON (Modeling)

DEFINITION: The teacher shows the students acceptable finished products or explains process before making an assignment.

SCALE DESCRIPTION: It is facilitating for students to not only know about, but to see examples of an acceptable finished product (story, poem, model, diagram, graph) or a process (how to identify main idea, weave). It is important that the visual input of modeling be critical elements of what is happening (or has happened) so students are focused on the essentials rather than being distracted by transitory or non-relevant factors in the process or product.

SCALE DESCRIPTORS:

A. Shows the students acceptable finished product or explains the critical elements of what is happening when giving an assignment.

B. Give an assignment without explanation or examples of what is expected.

C. Explains to the students acceptable finished product but does not show examples.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: EFFECTIVE LESSON (Checking for Understanding)

DEFINITION: The teacher checks for students' possession of essential skills to achieve the instructional objectives.

SCALE DESCRIPTION: The teacher needs to check for students' possession of essential information and also needs to observe students' performance to make sure they exhibit the skills necessary to achieve the instructional objectives. This can be done by:

a. Sampling: Posing questions to the total group in order to focus them on the problem and develop readiness to hear the answer, then getting answers from representative members of the group.

b. Signaled: Signaled responses from each of the total group. Selecting lst, 2nd, 3rd, 4th answer by showing that number of fingers, thumbs up or down for "agree" or "disagree", to the side for "not sure", raising hand when examples are correct.

c. Individual private response: Usually written or whispered to teacher so each student is accountable for demonstrating possession of, or progress toward achievement of the needed skills.

SCALE DESCRIPTORS:

A. Makes assignments after explanation of the instructional objectives.

B. Checks the students understanding

and information level by asking group questions, or using signaled or individual responses before making assignments.

C. Checks for understanding before making assignments by asking if there are any questions.

N. Not observed.

OBSERVATION SCALE

NAME OF SCALE: EFFECTIVE LESSON (Guided and Independent Practice)

DEFINITION: The teacher makes sure the instruction has taken before turning students' loose to practice independently.

SCALE DESCRIPTION: The beginning stages of learning are critical in the determination of future successful performance. Consequently, the students' initial attempts in new learning should be carefully guided so they are accurate and successful. Having instructed, teachers need to circulate among students to make sure the instruction has "taken" before the student can perform without major errors, discomfort or confusion, he/she is ready to develop fluency by practicing without the availability of the teacher. Only then students' can be given a written or verbal assignment to practice the new skill or process with little or no teacher direction.

SCALE DESCRIPTORS:

A. Circulates among students to make sure the instruction has taken before students practice independently.

B. Assigns independent practice without checking major errors, discomfort, confusion or circulating to help.

C. Circulates among students while they are practicing independently.

N. Not observed.